Assessing Vaccine Effects in HIV-1 Vaccine Trials: Antigenic Maps, Antigen Selection, and Sieve Analysis

Allan deCamp

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Program Authorized to Offer Degree: UW Biostatistics

University of Washington

Abstract

Assessing Vaccine Effects in HIV-1 Vaccine Trials: Antigenic Maps, Antigen Selection, and Sieve Analysis

Allan deCamp

Chair of the Supervisory Committee:
Professor Peter B Gilbert
UW Biostatistics

The goal of vaccination against infectious disease is a net population effect on the risk of infection and/or disease progression. In HIV-1 vaccine development efforts to date only a single HIV-1 vaccine trial has shown any efficacy by either of these measures. Despite this lack of success, we hope to inform future vaccine design by analyzing vaccine effects in HIV-1 vaccine trials. We expect to detect these effects through one of three measures: 1) differential vaccine-elicited responses among cases and controls in an immune correlates of risk analysis, 2) differential host genetics by treatment assignment among infected trial participants, or 3) differential viral genetics of breakthrough infecting strains by treatment assignment of the host (known as a sieve effect).

In this thesis we begin by developing methods for aggregating biomarkers for use in an immune correlates of risk analysis via antigenic maps. We show that antigenic maps can be used as a bridge to understand connections between immune correlates of risk analysis results, differential host genetics and local effects in a sieve analysis. Next, we discuss the critical step of identifying panels of antigens for use in bioassays that can be used to generate antigenic maps and compare immune response outcomes between HIV-1 vaccine trials. The last part of this thesis develops sieve analysis methods that focus on epitopes, the underlying unit of the adaptive immune system response.

TABLE OF CONTENTS

Page

List of Figures iv

List of Tables vi

Chapter 1: Introduction 1
1.1 A Brief History of HIV-1 Vaccine Development 2
1.1.1 The Step Study 5
1.1.2 RV144 HIV-1 Vaccine Efficacy Trial 6
1.1.3 Immune-Correlates of Risk 7
1.2 Research Goals 7
1.2.1 Antigenic Maps 8
1.2.2 Antigen Reagent Panel Selection 9
1.2.3 Sieve Analysis 9
1.3 Conclusion 10

Chapter 2: Antigenic Maps 12
2.1 Introduction 13
2.1.1 Theoretical Background 13
2.2 Hypothetical V2 Map 15
2.3 Generating a V2 antigenic Map 17
2.3.1 Statistical Considerations 17
2.3.2 V1V2-scaffolded Protein Antigenic Map 18
2.4 Biomarker combinations 18
2.5 HLA Antigenic Map Associations 20
2.6 Discussion 22

Chapter 3: Panel Selection 37
3.1 Introduction 38
3.1.1 Antigenic Space 39

3.2 V1V2-scaffold Protein Panel Selection . . . 41
3.2.1 Antigen assay characteristics . . . 42
3.2.2 Position of antigen within antigenic space . . . 42
3.2.3 Panel selection based on clusters . . . 43
3.2.4 Conclusions V1V2 panel selection . . . 45
3.3 Neutralization Serotype Discovery Project Panel Selection . . . 46
3.3.1 Magnitude breadth curves . . . 46
3.3.2 Statistical methods for panel selection . . . 47
3.3.3 Comparing methods of panel selection . . . 50
3.3.4 Cross validation results . . . 50
3.3.5 Predicting MB curves . . . 51
3.3.6 Panel Selection using grouped lasso . . . 51
3.3.7 Results of predicting MB curves . . . 52
3.3.8 Conclusion . . . 52

Chapter 4: K-mer Scanning . . . 65
4.1 Introduction . . . 66
4.2 Notation . . . 66
4.3 Statistical Methods: Hypothesis Testing and Estimation . . . 67
4.4 Simulation Methods . . . 69
4.4.1 Simulation Method 1 . . . 69
4.4.2 Simulation Method 2 . . . 73
4.5 Simulation Results . . . 76
4.5.1 Simulation Method 1 Results . . . 76
4.5.2 Simulation Method 2 Results . . . 76
4.6 Bias corrections to the sandwich variance estimator for small samples . . . 77
4.6.1 Generalized estimating equations and the sandwich estimator of variance 77
4.6.2 Simulation study of sandwich estimator bias corrections. . . . 81
4.7 Multiplicity Adjustment . . . 82
4.8 Discussion . . . 86
4.9 KC and mFG small sample adjustments are equivalent . . . 89

Chapter 5: Assessing Cytotoxic T Lymphocyte Effects . . . 101
5.1 Introduction . . . 102
5.2 Prior work . . . 103
5.3 Binding Scores . . . 104

5.3.1 Testing . . . 106
5.3.2 Alternative to using a "delta binding function" . . . 106
5.3.3 Testing . . . 108
5.4 Data simulation . . . 109
5.4.1 Data simulation results . . . 111
5.5 Analysis of Step trial . . . 112
5.5.1 Analysis of Step trial results . . . 113
5.6 Permutation analysis of PTE method . . . 115
5.6.1 Permutation tests . . . 117
5.6.2 Example using a two level HLA covariate . . . 118
5.7 Discussion . . . 124

Bibliography . . . 135

LIST OF FIGURES

Figure Number Page

2.1 Hypothetical Antigenic Map for V2 Binding Antibodies 24
2.2 Angles within the Hypothetical Antigenic Map for V2 Binding Antibodies . . 25
2.3 Shepard Plot for Antigenic Map of V1V2-scaffolded Proteins and RV144 Sera 26
2.4 Antigenic Map of V1V2-scaffolded Proteins and RV144 Sera 27
2.5 Antigenic Map of V1V2-scaffolded Proteins and RV144 Sera (Close Up) . . . 28
2.6 V2 Hotspot Amino Acid Distribution 29
2.7 Biomarker Combination Weights 30
2.8 Relationship between Spearman Correlation and Angles within the Antigenic Map 31
2.9 Antigenic Map Rotation 32
2.10 Location and HLA Odds Ratio Associations 33

3.1 Signal-to-noise versus P-value by dilution 53
3.2 Heatmap of Spearman Correlations, V1V2-scaffolded proteins baseline adjusted 1:100 dilution 54
3.3 Hypothetical V1V2 Antigenic Map 55
3.4 Heatmap of V1V2 Antigenic Map Angles 56
3.5 Optimized versus randomly selected virus panels 57
3.6 Tenfold Cross-Validation 58
3.7 Venn Diagram of Lasso Panels 59
3.8 Goodness-of-fit of Predicted MB-Curves 60

4.1 Cut Point and Probability Transformations 91
4.2 Probability-Probability Plot Assessing FWER 92
4.3 Sandwich Estimator Bias Corrections 93
4.4 Comparison of BH and BH_{ACT} under complete null hypothesis 94
4.5 Comparison of BH and BH_{ACT} under an alternative hypothesis 95

5.1 Simulation Study HLA frequency 125
5.2 Simulation study sequence mutation frequency 126
5.3 Power analysis: Linear PTE Model 3 127

5.4 Power analysis: Linear PTE Model 4 128

LIST OF TABLES

Table Number Page

2.1 V1V2-scaffolded Proteins . . . 34
2.2 Biomarker Combinations Correlates of Risk Results . . . 34
2.3 HLA Class I Antigenic Map Associations . . . 35
2.4 HLA Class II Antigenic Map Associations . . . 36

3.1 V1V2-scaffolded protein clusters . . . 61
3.2 Three Isolate Lasso Panel . . . 62
3.3 Nine Isolate Lasso Panel . . . 62
3.4 Three Isolate Tree Cluster Panel . . . 63
3.5 Nine Isolate Tree Cluster Panel . . . 63
3.6 Nine Isolate Panels Selected Using Lasso/Quartiles . . . 64

4.1 Size of test (MO, GEE, CWGEE) . . . 96
4.2 Power of test (MO, GEE, CWGEE) . . . 97
4.3 Coverage of test (MO, GEE, CWGEE) . . . 98
4.4 Power of test (MO,GEE, CWGEE) under informative cluster size alternative 99
4.5 Coverage of test (MO,GEE, CWGEE) under informative cluster size alternative 100

5.1 Simulation study HLA allele count frequency . . . 129
5.2 Simulation study sequence mutation statistics . . . 129
5.3 Simulation study power analysis . . . 130
5.4 Step Gag analysis . . . 130
5.5 Step analysis Gag 84 . . . 131
5.6 Analysis by HLA-A and HLA-B . . . 131
5.7 Analysis of most common HLA alleles . . . 132
5.8 Analysis of individual vaccine strain reference peptides . . . 132
5.9 Power analysis: Linear PTE Model 1 . . . 133
5.10 Power analysis: Linear PTE Model 2 . . . 133
5.11 Results of Step Permutation Analysis . . . 134

Chapter 1

INTRODUCTION

For many viral pathogens vaccination is a highly effective public health tool for controlling disease outbreaks. For example, smallpox has not only been controlled but it has been eradicated. Between 1967 and 1980, an intensive global eradication effort based on surveillance, containment, and vaccination successfully eliminated all natural cases of smallpox. A global polio eradication effort, based in part on vaccination, is ongoing; in 1988 polio was endemic to 125 countries while in 2012 it is endemic to only three. Influenza is another serious public health concern and at times infections have reached pandemic levels. When infected with the influenza virus most people recover; usually only the very young, elderly, and people with compromised immune systems are at risk of severe complications or death. However, during the 1918 influenza pandemic it was not the young and elderly who were most at risk of death but those aged 20 to 40. During this pandemic it is estimated that 27% of the world's population became infected with influenza and 3% of the world's population died as a result. Although there is a successful vaccine for influenza, annual vaccination is required to keep up with viral evolution and the threat of another global pandemic is very real. Other viral pathogens such as Hepatitis C, dengue, and HIV-1 continue to prove elusive in the effort to produce an efficacious vaccine. Recently however, an HIV-1 vaccine trial, RV144, demonstrated modest but statistically significant efficacy against acquisition of HIV-1 infection. Understanding patterns of the immune response induced by the vaccine used in RV144, the Step study, and the other HIV-1 vaccine efficacy trials to date, will likely provide insight into potential modifications of this and similar candidate vaccines that may lead to improved efficacy. Furthermore, methods developed for analyzing immune responses to vaccines for HIV-1 will likely translate to methods for analyzing immune responses to other viral pathogens.

1.1 A Brief History of HIV-1 Vaccine Development

Beginning in the 1980s researchers started to develope vaccines designed to prevent HIV-1 infection. Nearly all licensed vaccines against viral pathogens provide protection by generating serum or mucosa antibodies, typically to proteins exposed on the surface of the organism, that block infection or viremia [64]. Following this example, scientists began designing HIV-1 vaccines by engineering antigens based on the HIV-1 envelope glycoprotein

(gp120). The hope was that exposing the humoral immune system to this protein would allow the host to mount an antibody response that would be protective against infection. It became apparent starting in the early 1990s that this approach may not result in an effective vaccine. The potent responses generated against lab adapted strains of HIV-1 were not effective at blocking primary isolates. Researchers began to realize that it would be more difficult than first anticipated to succeed at this approach due to the high genetic variability seen in the HIV-1 envelope protein and the numerous ways that this virus shields itself from the immune system.

The first two HIV-1 vaccine efficacy trials, Vax004 and Vax003, were completed in 2002 and 2003 respectively. Vax004 which enrolled 5417 volunteers, most of whom were non-injection drug using men who have sex with men, was conducted in the United States, Canada, and the Netherlands and had a vaccine efficacy (VE) estimated at 6% (95% CI, -17% to 24%) [28, 16]. Vax003 which enrolled 2546 injection drug users was conducted in Bangkok, Thailand and VE was estimated at 0% (95% CI, -31% to 24%) [63]. Both vaccines were bivalent recombinant gp120 constructs. Both recombinant gp120 proteins in the Vax004 vaccine were based on subtype B strains matching the circulating strains in the study population. The Vax003 construct contained both a subtype B and a subtype E component matching the circulating strains in Thailand. The disappointing results of the Vax004 and Vax003 trials marked a turning point in HIV-1 research. The hope of achieving sterilizing immunity through an antibody based vaccine diminished and attempts to design a vaccine that might slow disease progression rather than prevent the establishment of infection were considered.

In an attempt to find another solution, researchers turned to vaccine strategies that would stimulate the cellular arm of the adaptive immune system. The cellular arm of the immune system can be effectively activated by all proteins of a pathogen; thus, vaccines can be developed that include relatively conserved proteins within the HIV-1 genome. Long-term non-progressor (LTNP) cohorts showed that in some cases strong Cytotoxic T Lymphocyte (CTL) responses were able to control disease progression and non-human primate (NHP) studies of vaccines designed to stimulate the cellular arm of the immune system were promising. Priming the immune system with HIV-1 proteins carried by a vector such

as adenovirus or poxvirus might invoke a response in the general population similar to that seen in the LTNP cohorts. Following the principle that cell-mediated immunity (CMI) may lead to a vaccine that does not protect via sterilizing immunity but might be effective at controlling HIV-1 replication in infected participants, two large efficacy trials referred to as the Step study and Phambili were conducted. The Merck vaccine used in these studies was designed to induce CMI to HIV-1 proteins Gag, Pol and Nef. The trials were designed to test the efficacy of the vaccine based on two outcomes; risk of infection and set point viral load. In 2007, immunizations were discontinued in the Step study after an independent Data Safety Monitoring Board (DSMB) reviewed results of a pre-defined interim safety and efficacy analyses and recommended that vaccinations be discontinued. This prompted the discontinuation of vaccination in the Phambili study as well. This failure marked a transition in the HIV-1 vaccine field back towards an interest in designing antibody based vaccines. Lessons learned in the build up to the big trials to date had lead people to understand that a two part approach that engages both the humoral and the cellular arms of the immune system may be best. It is well known that the two arms of the adaptive immune system are synergistic. As emphasized by Plotkin in a review of protective immune responses induced by vaccination,"multiple immune responses interact to protect" [64] and the field is aggressively pursuing vaccines designed to activate multiple components of the immune system including both arms of the adaptive immune system, CMI and humoral immunity, as well as innate immunity.

Meanwhile, advances in HIV-1 immunology were being made resulting in a better understanding of the structural elements of the envelope protein and targeted CTL epitopes. For example, the dynamic nature of the trimeric envelope spike [9] and the existence of an evolving glycan shield [83]. In addition, the discovery of many broadly neutralizing antibodies had been made which have been eclipsed by the discovery of even broader neutralizing antibodies [81, 80, 5]. Vaccine strategies had become more sophisticated and include using scaffolded proteins [38, 2, 27, 56], mosaic inserts [74], heterologous prime-boost inserts [40] and polypeptide targeted eptiope constructs [37]. Discovery of the tripeptide mimic of the $\alpha 4\beta 7$ integrin [1], the conserved CD4 binding loop, conserved immunological and structural features in the V2 and V3 loops [90], and the discovery of broadly neutralizing antibodies

found in natural infection all suggest that HIV-1 has vulnerabilities. Both arms of the adaptive immune system target epitopes - short fragments or small discontinuous regions of amino acids recognized by either T- or B-cell receptors. Much of this thesis will focus on the analysis of HIV-1 epitopes either directly via sequence analysis or indirectly via immune markers. In both cases we are interested in the variation in epitopic regions and the interplay between this variation and the host immune system; responses to these regions may have been vaccine induced or the result of natural infection. Another important topic of this thesis is the assessment of which antigen reagents should be selected for comparison studies of immunogenicity. Careful selection of reagents can optimize the information collected from immunogenicity studies and improve the power to detect relevant differences in immune responses.

As the field continued to evolve in the post Step study era, the RV144 HIV-1 vaccine efficacy trial, which employs a prime-boost strategy, was quietly progressing in Thailand. One of the components of this vaccine uses a canarypox vector designed to generate CMI and the other component was the bivalent subtype B/E recombinant gp120 vaccine used in the VAX003 trial designed to induce an antibody response. Results of the trial showing 31% vaccine efficacy were announced in September of 2009 which sparked great interest in understanding the mechanism by which the vaccine was protective. Given the modest efficacy of the vaccine there was a great interest in using the results of the RV144 trial to inform future iterations of HIV-1 vaccines.

Two of the efficacy trials mentioned above, the Step study and RV144, have a prominent role in this thesis and provide motivation and applications for the methods described in the following chapters.

1.1.1 The Step Study

The Step study was a phase II, randomized, double-blind, placebo-controlled test-of-concept study conducted at sites in North America, the Caribbean, South America, and Australia in 3000 HIV-1-seronegative volunteers. Participants were randomized to receive three injections of MRKAd5 HIV-1 gag/pol/nef vaccine (n=1494) or placebo (n=1506). Random-

ization was stratified by sex, four baseline levels of Ad5 antibody titer, and study site. The primary objective of the study was to assess reduction in HIV-1 acquisition or a decrease in HIV-1 viral load set-point. As part of the study design, futility boundaries were defined and at the first interim analysis for futility the study was stopped.

At the pre-specified interim analysis all but one infection had occurred in men. In participants with baseline Ad5 antibody titer less than or equal to 200, 3% of vaccine recipients (24 of 741) and 3% of placebo recipients (21 of 762) became HIV-1 infected (hazard ratio of 1.2, 95% CI 0.6-2.2) [8]. The geometric mean viral load was 4.61 log10 copies per mL in infected male vaccine recipients versus 4.41 in infected placebo recipients. In exploratory analyses, increased risk of acquisition of HIV-1 in the vaccine group compared to the placebo group was detected in two subgroups, Ad5 seropositive men (hazard ratio 2.3, 95% CI 1.2-4.3) and uncircumcised men (hazard ratio 3.8, 95% CI 1.5-9.3).

Follow-up studies largely focused on the issue of increased risk in the Ad5 seropositive and uncircumcised male subgroups. However, a study of breakthrough sequences from HIV-1 infected volunteers in the Step study provided evidence of selective pressure from vaccine-induced T-cell response on HIV-1 infection [71].

1.1.2 RV144 HIV-1 Vaccine Efficacy Trial

The RV144 trial was a phase III HIV-1 vaccine efficacy trial conducted in the Rayong and Chon Buri provinces of Thailand that enrolled 16,402 men and women primarily at risk for heterosexual infection. In a modified intention-to-treat analysis, vaccine efficacy was estimated at 31% (95% CI, 1% to 52%) [67].

After the results were announced a study of immune correlates of risk was undertaken to understand the potential mechanisms of protection. By November of 2009, pilot studies had begun in order to select assays for measuring vaccine-elicited immune responses that would be run in a case-control analysis. Using these pilot studies, a team of researchers identified 17 assay types based on each assay's ability to identify a vaccine induced response, the immunological function being tested and assay reproducibility. From the 17 assay types, six primary and 152 secondary readouts were analyzed in a case-control analysis in

which two of the six primary variables correlated significantly with HIV-1 infection risk [30]. Binding IgG antibodies to variable loops 1 and 2 (V1V2) of the HIV-1 envelope protein negatively correlated with infection and have been hypothesized to be a correlate of protection. The second marker, the binding of plasma IgA antibodies to the HIV-1 envelope protein, correlated positively with infection risk and it has been suggested that these responses may block the protective effects of other antibody responses. Given the modest efficacy of the vaccine, there is a need to use the results of this trial to inform future iterations of HIV-1 vaccines.

1.1.3 Immune-Correlates of Risk

The analysis above assessed immune-correlates of risk (CoRs) [66] by looking for associations between response biomarkers and infection rate among vaccinated subjects. These CoRs may not predict vaccine efficacy (VE) because they may simply correlate with natural (not vaccine-affected) susceptibility. A CoR may be hypothesized to predict VE which, if true, means it is a correlate of protection (CoP) [65, 66]. In the case of a CoP which is mechanistically and causally responsible for protection, we refer to it as a mechanistic CoP (mCoP) [65]. A CoR that is not a CoP is of interest since it may provide insight into the immunology of HIV-1 susceptibility which could lead to better vaccine designs. A CoP that is not an mCoP is also of interest in that it can potentially be used to bridge an effective vaccine between populations. If we do identify an mCoP we stand the best chance to make incremental modifications to the vaccine that will improve efficacy. This dissertation describes methods for testing and assessing CoRs, CoPs, and mCoPs, together referred to as immune correlates. Our methods are based on immune biomarker responses and sequence data from infecting viral strains and our goal is to guide future vaccine development.

1.2 Research Goals

The goal of this dissertation is to describe methods and analyses designed to understand patterns of vaccine induced immune response which, when applied in the context of an efficacy trial, assess the impact of these responses on vaccine efficacy. This dissertation

focuses on HIV-1 vaccine development but the methods can be applied more generally to vaccine development for infectious diseases in general.

1.2.1 Antigenic Maps

Chapter 2 of this dissertation will compare methods that combine multiple readouts from a set of related antigens into a single biomarker for evaluation as an immune correlate. Any hierarchical clustering tree of antigens can be used to define a set of antigen weights using the Gerstein-Sonnhammer-Chothia method [19] which corrects for unequal representation within the tree. This set of weights can be used to combine the multiple antigen readouts into a single biomarker. Several different weighting schemes will be considered: 1) equal weighting, 2) weighting based on a phylogenetic tree, 3) weighting based on pairwise correlations of antigen readouts, and 4) weighting based on an antigenic map. How to optimally assign weights based on an antigenic map is unknown and to our knowledge has not been studied. One goal of this chapter will be to address how to assign weights based on the position of antigens within an antigenic map. To assess the performance of these weighting schemes, we will consider the scenario where there is an underlying antigenic map of a specific receptor/ligand interaction that defines the magnitude of the assay readouts to multiple cross-reactive antigens. It will also be assumed that these readouts can be combined to form a single biomarker readout that measures the coverage of the vaccine induced response with respect to circulating strains. Using RV144 case-control samples, we will compare the results of an immune correlates analysis of multiple V1V2-scaffolded proteins using the weighting schemes described above which may generate a hypothesis for which type of immune correlate has been discovered (i.e. CoR, CoP, or mCoP). Additionally, assessment of immune correlates in the context of an antigenic map, where antigens constructed from breakthrough infections have been measured, may lead to the direct assessment of VE as a function of location within the antigenic map.

1.2.2 Antigen Reagent Panel Selection

Chapter 3 will cover methods for identifying antigen panels, sets of reference reagents, for the evaluation of vaccine-elicited immune responses. The results of the RV144 immune correlates study [30] highlight the need for well defined antigen panels that can be used for current and future studies. Common antigen panels can be used to compare results across studies and can help inform which vaccine candidates to advance to efficacy trials. They may also help identify modifications to current vaccines that would improve vaccine efficacy.

In this chapter we will provide two examples of antigen panel selection. The first example describes the method used to down select a set of V1V2-scaffolded protein reagents that ultimately became the set of antigens used in the antigenic map described in the previous chapter. This set of reagents has also been used to inform decisions regarding future efficacy trials based on the analysis of the V1V2 responses to current vaccines. The second example describes the selection of a panel of viral isolates for use in neutralization assays. A vaccine that can elicit neutralizing antibodies that protect against a broad range of globally circulating strains has been termed the "Holy Grail" of HIV-1 vaccine research [14]. In order to assess the neutralization capabilities of candidate vaccines, a small panel of isolates that can predict the ability of sera to neutralize a broad range of strains is needed. In our second example we describe methods for selecting such a panel and describe the predictive ability of these panels.

1.2.3 Sieve Analysis

Chapters 4 and 5 cover sieve analysis methods and applications. Vaccine sieve effects are vaccine induced adaptive immune responses that partially block HIV-1 acquisition and/or alter HIV-1's evolution post-acquisition. Sieve analysis can shed light on potential mechanisms of immune protection, as well as generate hypotheses on which immunological measurements may be immune correlates of protection.

Dating back to 1998 Gilbert and colleagues have been working on sieve analysis methods (e.g., [23, 20, 22, 24]). These methods either focus on a single sequence per subject or focus on scanning one site at a time. Manuscripts by deCamp with Rolland et al. 2011,

Gnanakaran et al. 2011, Janes et al. 2012, and Rolland et al. 2012 [71, 25, 36, 70] use novel methods for assessing multiple sequences per subject and collections of sites but the methodology has never been completely described and studied.

Chapters 4 and 5 improve and expand upon the methods developed for those applied analyses. A set scanning method is used in [71] that needs further development and characterization and is addressed in Chapter 4. This method uses a sliding window where the size of the window is typically 9 amino acids, a common T-cell epitope length. Since the windows overlap by 8 of 9 amino acids and multiple epitopes are analyzed, methods for multiplicity correction are needed to obtain FDR-adjusted and FWER-adjusted p-values. Host genetics may be critical in understanding and fully characterizing a vaccine effect so methods that can incorporate covariates are of interest. Of particular interest is the incorporation of HLA genotypes into the analysis of sequence data which creates a unique challenge for methods development due to the diversity of HLA genotypes. Sieve methods have been used that incorporate predicted T-cell epitopes based on subject HLA typing [71] and pre-infection T-cell responses as measured by ELISpot [36]. T-cell epitope prediction is an active area of research and expansion and development of methods that incorporate predicted T-cell responses are needed. In Chapter 5 we expand the method described in Chapter 4 to handle host HLA genotype information.

1.3 Conclusion

Bold claims that a vaccine for HIV-1 would be developed within two years were made when the virus identified as the cause of AIDS was first identified. Since then 30 years of vaccine development have resulted in one modestly successful vaccine. Whether or not that result can be replicated in the original population and expanded to new populations has yet to be determined and will not be known for many years. In this dissertation the aim is to identify methods for assessing vaccine effects in HIV-1 vaccine trials. Through antigenic maps, reagent selection and sieve analysis we hope to help bridge the gap between rational vaccine design and empirical based vaccine development.

The current effort to rationally design an HIV-1 vaccine needs to be pursued,

> although there are voices that warn that a strict reductionist approach may actually hamper the effort. Rational vaccine design efforts need to be complemented with strategically planned empirical immunogenicity and efficacy evaluations [14].

In the quote above, Esparza captures the tension between two current thoughts on how best to achieve an effective HIV-1 vaccine. However, the idea that one means of achieving a vaccine is better than the other is really a false dichotomy. Empirical vaccine design is rational and rational vaccine design cannot exist without empirical results. The current trend in HIV-1 vaccine design appears to be towards quicker, smaller trials, with an iterative approach that will allow empirical results to facilitate the design of the next generation of vaccines. This is exactly the approach I advocate and the methods developed in this thesis are intended to inform a rational, iterative approach to incrementally improve existing vaccines which have historically, and perhaps necessarily, been based on empirical design.

Chapter 2

ANTIGENIC MAPS

2.1 Introduction

Antigenic maps have a long history of theoretical and applied research in influenza vaccine development. The World Health Organization (WHO) coordinates the development of influenza vaccines using a worldwide surveillance network and bases decisions on vaccine design in part based on antigenic maps generated from real time data collection [57, 18]. In the field of HIV-1 vaccine research this method for designing and assessing vaccines has not be employed. This is in part due to the lack of an effective vaccine and also because the data to support the generation of antigenic maps for HIV-1 have not been widely available. In influenza, the hemagglutination inhibition (HI) assay measures the ability of antibodies to block the binding of the influenza hemagglutinin (HA) envelope protein to sialic acid receptors on host cells; the HI assay measures antibodies that disrupt or prevent a specific receptor/ligand interaction. We are not aware of assay data that have been successfully used to generate an HIV-1 antigenic map to date other than the map presented in this chapter. The available data appear to be in many cases too coarse in that HIV-1 assay readouts often measure receptor/ligand interactions to multiple antigenic determinants simultaneously. With the discovery of binding antibodies to the V1V2 region of HIV-1 Env as a correlate of risk in the RV144 trial, and further studies which suggest that the RV144 vaccine induces a V2 response targeting β-strand C on the N-terminal side of the putative $\alpha 4\beta 7$ binding motif [12, 89], we expect that HIV-1 vaccine research can benefit from the use of antigenic maps specific to this region. Furthermore, given targeted assays measuring antibody responses to multiple strains in a region of interest (e.g., V3 loop, C1 region, CD4 binding site, bridging sheet, etc.) we expect antigenic mapping can inform HIV-1 vaccine design.

2.1.1 Theoretical Background

In 1979 Perelson and Oster [60] introduced the concept of shape space which suggests that antibodies and antigens can be co-embedded in a low dimensional Euclidean space. Viral antigens and host antibodies are complementary in that when they bind they form a lock and key structure of varying affinity. The stronger the interaction, as measured by a given assay,

the closer the antigen and antibody are in shape space. By comparing the location of vaccine induced antibodies, the vaccine strain, and circulating strains, an antigenic map suggests an iterative approach to vaccine design in which the placement of one ore more vaccine strains within the map can improve the coverage of the vaccine induced antibodies relative to circulating strains and thereby improve vaccine efficacy. The theoretical framework of antigenic maps was applied by Lapedes and Farber to influenza in 2001 [42] and in 2004 Smith et al. compared antigenic and genetic evolution in H3N2 from 1968 to 2003 and suggested that the methodology is readily applicable to monitoring and estimating vaccine efficacy in influenza [76].

The generation of an antigenic map requires a data matrix of assay readouts from n serum samples against m antigens. A version of multidimensional scaling (MDS) known as multidimensional unfolding (MDU) is applied to the data which embeds each serum sample and antigen in the same low-dimensional space. One goal of this chapter is to compare methods for combining the readouts from multiple antigens into a single biomarker where one of these methods will combine the antigens using the location of each antigen within an antigenic map. Alternative methods for combining readouts include using an equally weighted average which is appealing because it is simple but does not incorporate any information regarding the relatedness of the antigens. Another approach for combining the antigens into a single biomarker is to combine the assay readouts of the m antigens based on a phylogenetic comparison. We know from Smith et al. [76] that although phylogeny and antigenicity tend to track each other, small changes in phylogeny can produce large changes in antigenicity and vice versa. This suggests that a method for combining antigens into a single biomarker based on antigenic maps will be a more powerful approach than using phylogeny when assessing immune correlates. The primary IgA variable in the RV144 immune correlates study [30] was a weighted average of 14 antigens using the Spearman correlations between the antigen readouts. This method uses the antigenic properties of the readouts to create a score without directly using an antigenic map. In the case of assessing binding antibodies to the region of HIV-1 it is expected that computing weights directly from an antigenic map will lead to a stronger immune correlate than the other weighting options just mentioned. In this chapter we generate an antigenic map based on the peak

immunogenicity readouts for IgG binding antibodies as measured by the binding antibody multiplex (BAM) assay from the 246 RV144 case-control sera. This is a step in the direction of our ultimate goal which is to understand VE as a function of the location of the response within an antigenic map which, as already mentioned, might suggest modifications of the vaccine strain that will lead to improved vaccine efficacy.

In the sections below we infer that our antigenic map is V2 specific based on epitope mapping experiments which suggest that the IgG binding antibody response to V1V2 antigens are directed towards $\beta-$strand C of the V2 loop. In this chapter we examine a hypothetical V2 antigenic map in Section 2.2, we create an antigenic map for V2 binding antibodies in Section 2.3, we assess biomarker combinations in Section 2.4, we consider HLA genotype associations with antigenic map location in Section 2.5, and we finish the chapter with a discussion of antigenic maps in Section 2.6.

2.2 Hypothetical V2 Map

Variable loop 2 (V2) is one of 5 variable loops on the HIV-1 gp120 protein. The name is somewhat misleading because elements of this and other variable loops are quite conserved. For example, the N-terminal portion of V2 is consistently 28 amino acids in length, has low entropy at these positions, typically contains an N-linked glycosylation site at position 160 and usually contains an $\alpha 4\beta 7$ binding motif on the C-terminal side of the loop crown region. This region was found to be highly antigenic for the RV144 vaccine and several assays were used to identify antibody activity directed towards this region. Furthermore, responses at this location were found to be inversely correlated with infection [30]. This generates the hypothesis that antibodies directed towards this region may be protective.

A detailed antigenic map may be useful in the generation of future vaccines by allowing the placement of vaccine antigens at key positions within the map. Figure 2.1 is a hypothetical map based on data that could have been be collected in the RV144 trial. In this figure we assume a 2-dimensional map generated using BAM assay readouts where V2 binding measurements were collected for the 3 vaccine strain antigens along with a sampling of circulating strains. The ideal data generation would include antigens fabricated from the sequences of breakthrough infections in order to truly represent circulating strains. In anti-

genic cartography we co-embed in a map the sera responses (in this hypothetical example we consider V2 monoclonal antibodies purified from vaccine recipient sera) with these antigenic targets. Sera responses are categorized as either High or Low/Medium with respect to the binding response to one of the antigens (in this case the MN vaccine strain antigen). In the map the distance between the antigen and sera is a monotonically decreasing function of the estimated assay readout; the monotonic relationship is between the estimated readout and the distance on the map not the true readout due to the dimension reduction step of MDU which is discussed in more detail in Section 2.3.1.

If some functional aspect of V2 binding antibodies are in fact protective, we would expect the highest binders to overlap most closely with the circulating strains since distance within the map relates directly to the tightness of binding as measured by the BAM assay. In our hypothetical map we would expect binding to MN to be an inverse correlate of risk while the other two vaccine strain antigens would not correlate with risk since high binding to MN implies a sera response close to the circulating strains while high binding to either of the other vaccine strains does not indicate the location of the sera with respect to the circulating strains. However, the antigenic map allows us to explore measurements other than the assay readout (or map distance) to detect correlates of risk. For example, within the map if we use 92TH023 as the vertex, then the angle between the lines connecting 92TH023 to the center of the circulating strains and a sera response would correlated with risk since sera with angles close to 0 are closer to the circulating strains than sera with angles close to π radians (see Figure 2.2). The $x-$ and $y-$coordinates together should also correlate with risk. Rotating the map in Figure 2.1 by $\pi/2$ radians[1] would make the $x-$coordinate a correlate of risk while the odds ratio for the $y-$coordinate would be approximately 1. From this exercise we hypothesize that the map coordinates and angles between sera responses, antigens, and the center of circulating strains are the most relevant measures for testing for correlates of protection. We also hypothesize that rotating the map may isolate the association with risk to one of the coordinates.

[1]Positive rotations are in the counterclockwise direction, negative rotations are clockwise.

2.3 Generating a V2 antigenic Map

In response to the finding that IgG binding antibodies to a V1V2-scaffolded antigen [30] inversely correlated with infection risk, follow-up assays were done with additional V1V2-scaffolded antigens. The original correlate was based on an ELISA assay using the clade B HIV-1 strain Case A2 scaffolded on a Murine Leukemia Virus (MuLV) gp70 envelope protein. As described in Chapter 3, additional studies were conducted using multiple V1V2-scaffolded antigens in both the BAM and the original ELISA assays. For the generation of an antigenic map we used the 246 week 26 case-control sera readouts to nine gp70 V1V2-scaffolded proteins. These proteins are listed in Table 2.1.

In order to generate an antigenic map, we first must convert the assay readout, in this case mean fluorescence intensity (MFI), into a dissimilarity measure using a monotonically decreasing function of the assay readout. For readout x, we use the transformation

$$d = 10.46 - log(x) \tag{2.1}$$

to derive the dissimilarity d where 10.46 was chosen by taking the maximum log(MFI) readout across all antigens and adding 0.1. The transformation function requires some thought and a thorough discussion of the choice of transformation function is beyond the scope of this chapter. For this case we chose a value which is very near the saturation point of the assay, 10.46, and subtract the log(MFI) readout. Additionally, working on the log scale rather than the raw assay readout is typical for creation of antigenic maps. The second choice that must be made when generating the map is the number of dimensions. We choose two dimensions which simplifies plotting and interpretation of the map. Optimization of the number of dimensions was not considered and is also beyond the scope of this chapter.

2.3.1 Statistical Considerations

Here we describe the algorithm used to generate an antigenic map using multidimensional unfolding (MDU). The goal is to approximate the transformed assay readout matrix $\{Y_{ij}\}$ consisting of n_1 sera by n_2 antigens using configurations X_1 and X_2 ($n_1 \times p$ and $n_2 \times p$ respectively where p is the dimension of the map). By transformed assay readout matrix we

mean after the monotonic transformation discussed above has been applied (see Equation 2.1). The configurations X_1 and X_2 are the coordinates of the sera and antigens respectively within the antigenic map. We have

$$Y_{ij} \doteq d_{ij}(X_1, X_2) \equiv \sqrt{\sum_{s=1}^{p}(X_{1is} - X_{2js})^2} \tag{2.2}$$

and we are interested in minimizing the stress function

$$\sigma(X_1, X_2) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{ij}\left(Y_{ij} - d_{ij}(X_1, X_2)\right)^2 \tag{2.3}$$

where $\{w_{ij}\}$ is a matrix of $n_1 \times n_2$ weights. We use the implementation of MDU found in the R library smacof (Scaling by MAjorizing a COmplicated Function) [44].

2.3.2 V1V2-scaffolded Protein Antigenic Map

Figure 2.3 shows the assay based dissimilarities as computed by Equation 2.1 versus the configuration distances based on the antigenic map which is known as a Shepard plot. A Shepard plot assesses the fit of our antigenic map to the data. The antigenic map is shown in Figures 2.4 and 2.5 where Figure 2.5 is a close up view to better view the relative locations of the nine V1V2-scaffolded proteins. In these two figures, the contour lines are generated by running a correlates analysis over a grid of hypothetical antigens. Each point on the map is the location of a potential V1V2-scaffolded protein. By taking the distances from a grid point to each of the 246 sera locations we can back transform to the a hypothetical assay readout and run a correlates of risk analysis as described in [30] with the addition of IgA as a covariate. For each grid point we compute the theoretical odds ratio at that point and display these values using contour lines.

In Figure 2.6, we plot the amino acid of each of the nine V1V2-scaffolded proteins for HXB2 positions 165 to 178; these amino acid positions are located within a region of V2 known as the "V2 hotspot" discussed in [30, 70].

2.4 Biomarker combinations

A primary goal of this chapter is to use an antigenic map to generate a combined biomarker readout and compare the results to other methods of combining the nine V1V2-scaffolded

protein readouts. As outlined in Section 2.1, we compare weighted averages based on 1) one minus the Spearman correlations between antigen readouts, 2) Phylogenetic tree distances between antigens, and 3) simply averaging the readouts to biomarkers derived from the antigenic map. As we did for the contour lines we run a correlates of risk analysis as described in [30] with the addition of IgA as a covariate for each of the three combined readouts. In addition to the combined readouts outlined above we compute a fourth readout based on the antigenic map using the pairwise angles between antigens within the map with a vertex defined by the medoid of the sera locations. $K-$medoids clustering is similar to the $K-$means clustering [41]. A medoid is an element of a collection of objects which minimizes the mean dissimilarity between objects. The medoid of the sera is the location of the serum selected by the 1-medoid clustering algorithm which minimizes the average Euclidean distance between that serum location and all other sera locations. In all of these cases, except the simple average, we compute a dissimilarity matrix for pairs of antigens and use the Gerstein-Sonnhammer-Chothia (GSC) method [19] for weighting based on this matrix. Once we have computed the GSC weights we compute the combined biomarker readout using those weights. The weights for each of the four combined biomarkers are shown in Figure 2.7.

In Table 2.2 we see that the correlates of risk results for the combined antigens are nearly identical for all four weighting schemes. Figure 2.8 shows that the pairwise Spearman correlations between antigen readouts and the pairwise angles computed from the map track very closely.

In addition to using angles within the map to generate weights for a combined biomarker we are also interested in analyzing the $x-$ and $y-$coordinates both separately and together as another way to combine the biomarkers. Again we apply the correlates of risk analysis of Haynes et al. [30] however this time we include both the $x-$ and $y-$coordinate in the model rather than a single biomarker in the first four examples. As before we include IgA as a covariate in our model of infection risk. We are particularly interested in how rotation of the map, as described earlier in our hypothetical analysis in Section 2.2, may isolate the correlation with risk to a single coordinate. To this end, for a sequence of 100 angles from 0 to 2π radians, we repeat the analysis after rotating the antigenic map through the specified

angle. We are not looking for the best model among 100 separate models in this analysis. Rather, the model is fundamentally the same with each rotation and we are simply looking for the rotation that best separates the effects of the $x-$ and $y-$coordinates. In Figure 2.9, we see that rotating the map approximately $^{7\pi}/_{4}$ radians isolates the effect of the $x-$ and $y-$coordinates in that $x-$coordinate is a strong inverse predictor of risk with an odds ratio less than 0.55 per 1 SD change in the coordinate whereas the $y-$coordinate is not a significant predictor of risk with an odds ratio of close to one per 1 SD change in the $y-$coordinate. The estimated odds ratio per 1 SD in the $x-$coordinate reaches a minimum for a rotation of 5.64 radians (odds ratio for $x-$coordinate 0.53, P-value 0.00047; odds ratio for $y-$coordinate 0.89, P-value 0.63).

2.5 HLA Antigenic Map Associations

We hypothesize that location within the antigenic map may be associated with HLA genotype. If such associations exist they should be restricted to HLA Class II alleles based on our understanding of Class II help of the antibody response. In order to have enough power to detect an association between HLA genotype and map location we will likely need to test alleles that are highly represented within the trial population. If weak associations between HLA alleles and map location exist we also hypothesize that these associations might be reflected in the allele frequencies of cases as compared to controls. That is, alleles that are associated with a protective location within the map might also be seen less frequently in the cases than the controls. Similarly alleles that are associated with a location that positively correlates with infection risk will likely be enriched in the cases. This enrichment, if seen, does not imply enhancement within participants with this allele because enrichment in one allele may simply reflect depletion of one or more protective alleles.

We first rotated the map 5.64 radians, as discussed in Section 2.4, in order to isolate most of the association between infection risk and location to the $x-$coordinate. We considered all 2-digit HLA Class I and Class II genotypes that are represented in at least 10% of the 246 vaccine recipients from the RV144 correlates analysis. We identified 17 Class I alleles and 19 Class II alleles. We then ran a likelihood ratio test for an association between HLA and map location by comparing two logistic regression models. The full model predicts HLA allele by

location ($x-$ and $y-$coordinate) in the map and sex. The base model predicts HLA allele using only sex. We also test for significant departures from one in the odds ratios associated with the individual $x-$ and $y-$coordinates using a Wald test. In addition to testing for significant associations between HLA allele and location we test for an association between HLA allele and infection status using a contingency table of counts of the presence versus absence of the allele given case or control status. We test this association using a Fisher's exact test and compute the estimated odds ratio using the conditional MLE.

Tables 2.3 and 2.4 show the results of these association tests for HLA Class I and HLA Class II alleles, respectively. As shown in Table 2.3, we find no significant associations between HLA Class I allele and map location; the p-values for the overall test of association appear to be nearly uniformly distributed. Seven of the 19 HLA Class II alleles tested have an unadjusted p-value less than 0.10 and two of these have a Holm adjusted p-value of less then 0.10. The strongest association between HLA allele and map location is with the DPB1:13 allele (unadjusted p-value 0.0001; adjusted p-value 0.001). The odds ratio per 1 SD in $x-$coordinate for this allele is estimated to be 2.02 (p-value 0.0003) and the odds ratio per 1 SD in $y-$coordinate for this allele is estimated to be 0.64 (p-value 0.10).

The tests for association between allele frequency and infection status showed no significant differences in allele frequency by infection status; p-values are nearly uniformly distributed within both the HLA Class I and HLA Class II alleles (see Tables 2.3 and 2.4). Despite the lack of significant associations we are still interested in trends between the allele frequency associations and the allele map location associations. For example, when we look at the direction of the effect with respect to the $x-$coordinate when testing DPB1:13, the association suggests an increased prevalence of DPB1:13 with increasing x. Together with the overall correlation between the $x-$coordinate and infection risk, these two associations suggest that we would see a lower frequency of DPB1:13 in the infected than the uninfected groups which in fact is the trend in the data (11 of 41 cases are DPB1:13 versus 77 of 205 controls; odds ratio 0.61; see Table 2.4). Similarly for an allele with an odds ratio of less than one for an association with the $x-$coordinate, we might expect to see an increased frequency of that allele in the cases as compared to the controls. For example, DRB1:07 is associated with the $x-$coordinate with an estimated odds ratio of 0.5 and this allele is seen

in 11 of 41 cases and 31 of 205 controls (odds ratio 2.05; see Table 2.4).

We explore these associations for all HLA Class I and HLA Class II alleles for both coordinates in Figure 2.10. We make a scatter plot of the odds ratio for association with infection across HLA Class I (or HLA Class II) alleles against the odds ratio for association with $x-$coordinate (or $y-$coordinate) with those same alleles. We see in Figure 2.10 that the strongest trend is between odds ratios for the Class II allele frequencies and the $x-$coordinate associations. This trend is in the direction we would expect as previously described using the two examples based on the alleles with the strongest associations (DPB1:13 and DRB1:07).

2.6 Discussion

In this chapter we have demonstrated that an antigenic map of a specific IgG binding antibody response to HIV-1 antigens is possible. Furthermore this map is useful both for generating a combined biomarker based on the responses to multiple antigens as well as using the coordinates of the map directly as biomarkers in a correlates of risk analysis. With the map we are also able to assess associations between host genetics (e.g. HLA Class II allele) and coordinates within the map. We can find biologically plausible trends that connect host genetic factors to infection risk as a function of the specific immune response captured by the map (analysis presented in Figure 2.10). The map also appears to capture the sequence space of the antigens and this should be explored further (see Figure 2.6); notice that at every position within the V2 hotspot region antigens with the same amino acid signature at that position cluster within the map.

Clearly there are shortcomings to this antigenic map. For example, we have only one of the three vaccine insert constructs represented. Also, the number of antigens used to generate the map is quite small with only nine antigens and many of these are either identical in sequence (e.g. case A2 antigens generated in three different labs) or very similar in sequence (e.g. the V169K variant of case A2 and the "mut3" variant of case A2 which is sequence identical to the AE antigen at the putative binding site). We would also like to see antigens that are based on breakthrough infections as these would likely expand our ability to interpret the map. The data needed to generate these antigens are available since we know the V1V2 sequences for many of the breakthrough infections. The only barrier to creating

these antigens and the associated map is the time and cost associated with developing these antigens in the laboratory and communicating the utility of readouts to these antigens. As a direct example of the utility of antigens derived from breakthrough infections, we cite the follow up of the Vax004 trial. In Gilbert et al. [21], the authors demonstrate that the neutralization potency of vaccine-elicited antibodies against early viruses from placebo recipients was stronger as compared to early viruses from vaccine recipients.

Perhaps the greatest utility of a detailed antigenic map would be the ability to hypothesize which virus strains should be used within a vaccine. If IgG binding to V2 is a mechanistic correlate of protection then the map suggests the next iteration of this vaccine should shift the V2 response to the right along the $x-$axis of the map after rotating 5.64 radians. Here the utility of knowing the location of the circulating strains is paramount. The analysis in this chapter suggests that if we had antigens based on circulating strains they would in general be located within the map to the right of the observed IgG binding responses but we do not know by how much. It seems possible but implausible that the current vaccine strains provide optimal coverage for the circulating strains seen in this trial. More likely, changing the vaccine inserts in such a way that would shift the binding responses to the right without overshooting the mark would improve coverage and hence the effectiveness of the vaccine. Additional early phase trials could be run which would modify the existing vaccine in ways that would shift the V2 binding antibody response and comparisons of coverage between vaccine regimens could be directly tested with the results providing direct implications to vaccine efficacy.

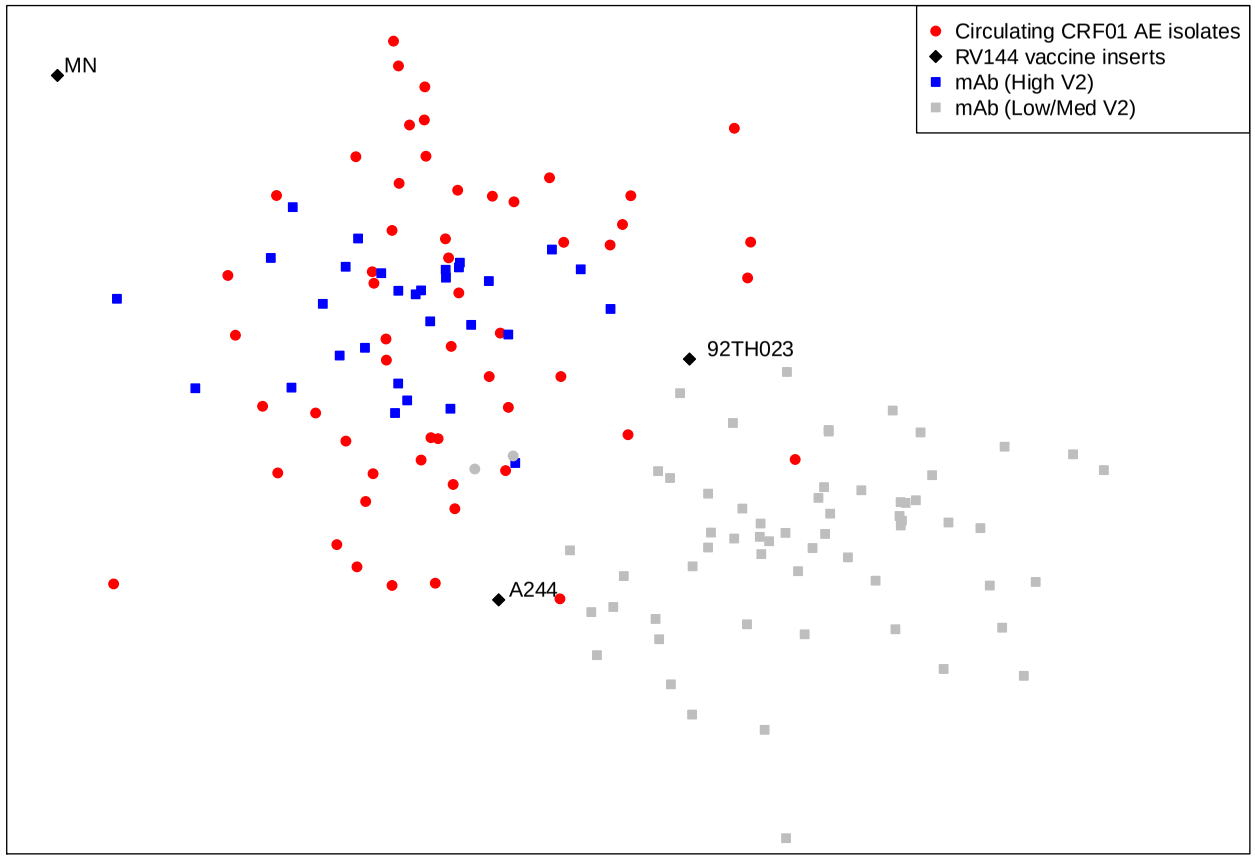

Figure 2.1: Hypothetical antigenic map of V2 binding antibodies elicited by the RV144 vaccine. Red dots represent the location of antigens derived from early viruses collected from breakthrough infections during the trial. The black diamonds are the locations of the three vaccine inserts. Blue (and gray) squares show the location of V2 specific monoclonal antibodies purified from vaccine recipient sera with High (and Low/Medium) response to the MN vaccine strain.

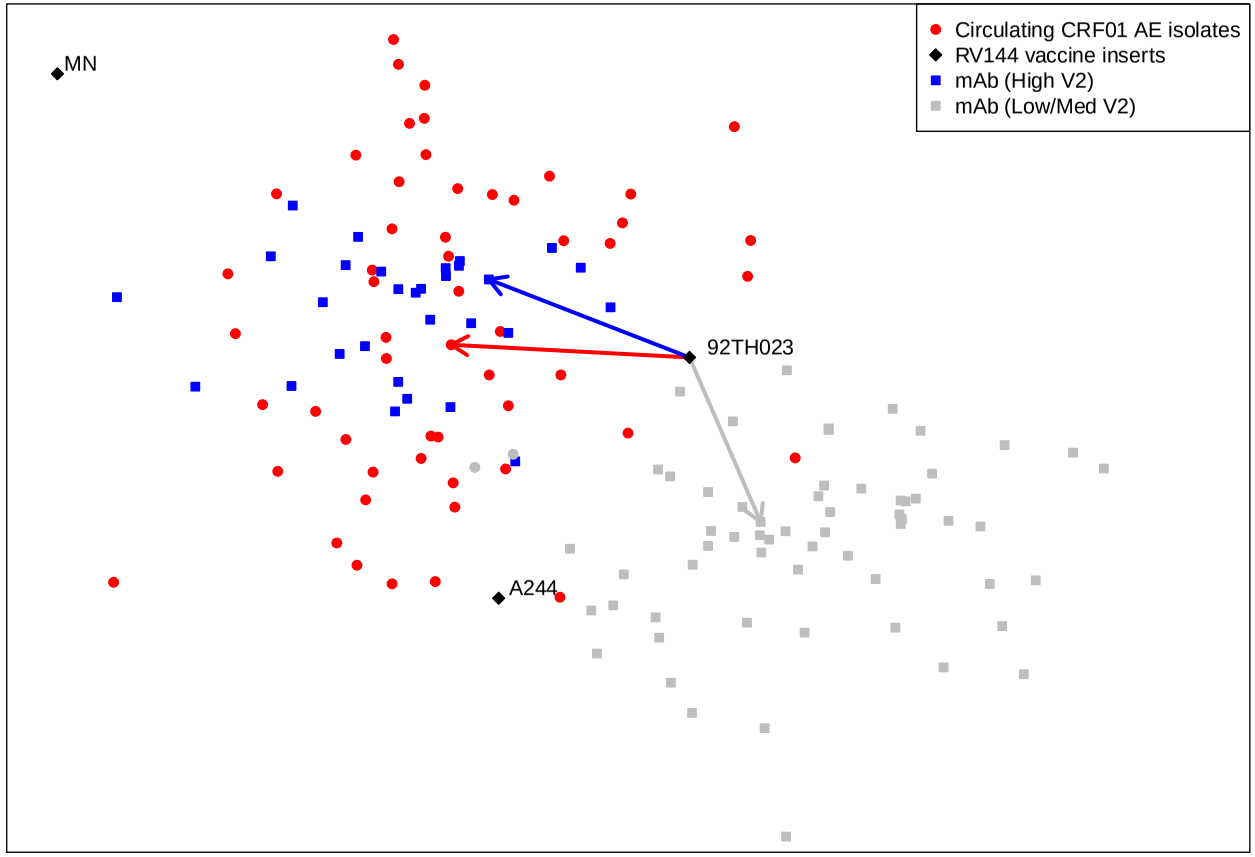

Figure 2.2: Angles within the hypothetical antigenic map of V2 binding antibodies elicited by the RV144 vaccine. Points within the map are defined in the caption of Figure 2.1. Two angles are shown, both with vertex at the location of 92TH023 within the map. One side of each angle, shown in red, is defined by the segment connecting 92TH023 to the medoid of the circulating isolates (see Section 2.4 for the definition of medoid). The second side of the first angle, shown in blue, is between 92TH023 and a randomly selected High binding mAb. The second side of the second angle, shown in gray, is between 92TH023 and a randomly selected Low/Medium binding mAb. The two angles measure -0.44 and 1.94 radians, respectively demonstrating that the angle between the medoid of the circulating strains and the High binding mAbs with vertex at 92TH023 tends to be close to zero while the corresponding angles defined by Low/Medium binding mAbs tends to be close to π.

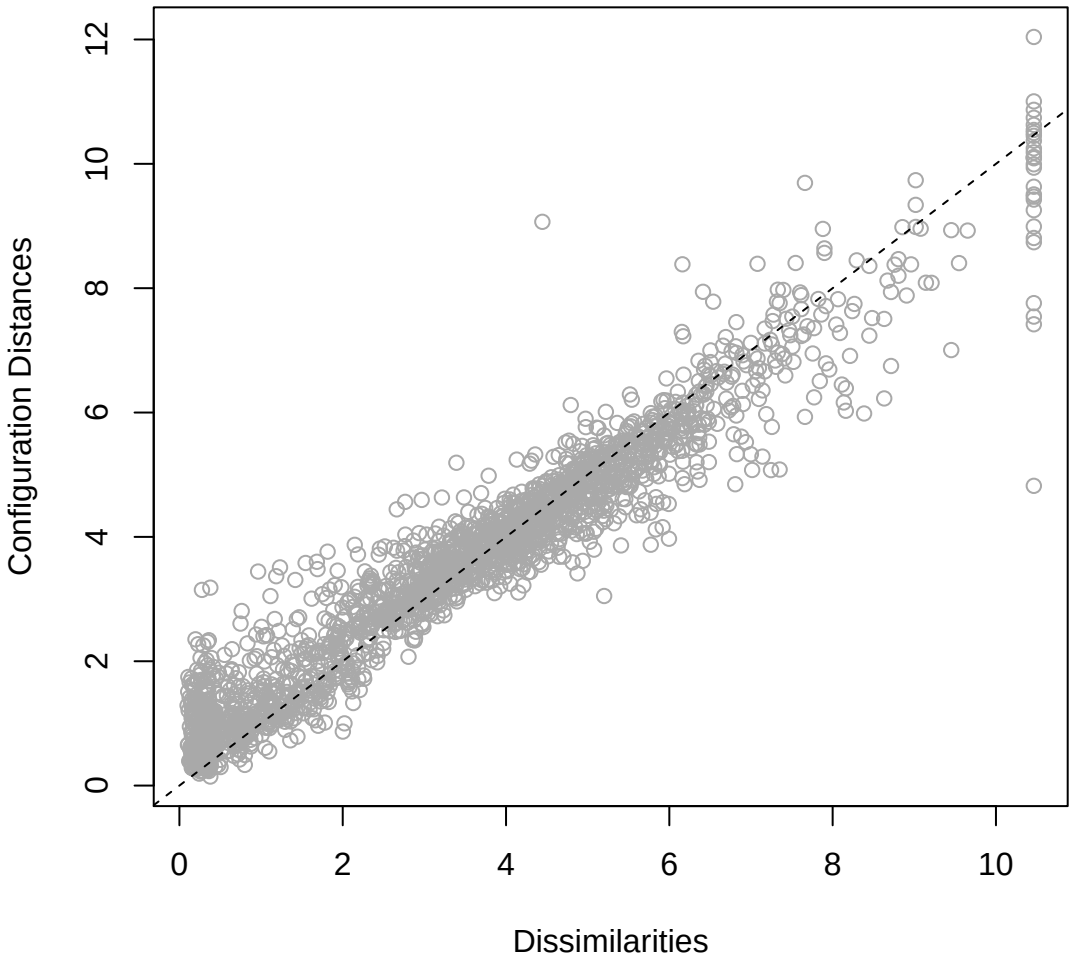

Figure 2.3: Shepard plot for the antigenic map of nine V1V2-scaffolded proteins assayed in the BAM assay against the 246 RV144 case control sera. A Shepard plot is a scatter plot of the observed dissimilarities against the fitted distances. The dashed line is the identity line.

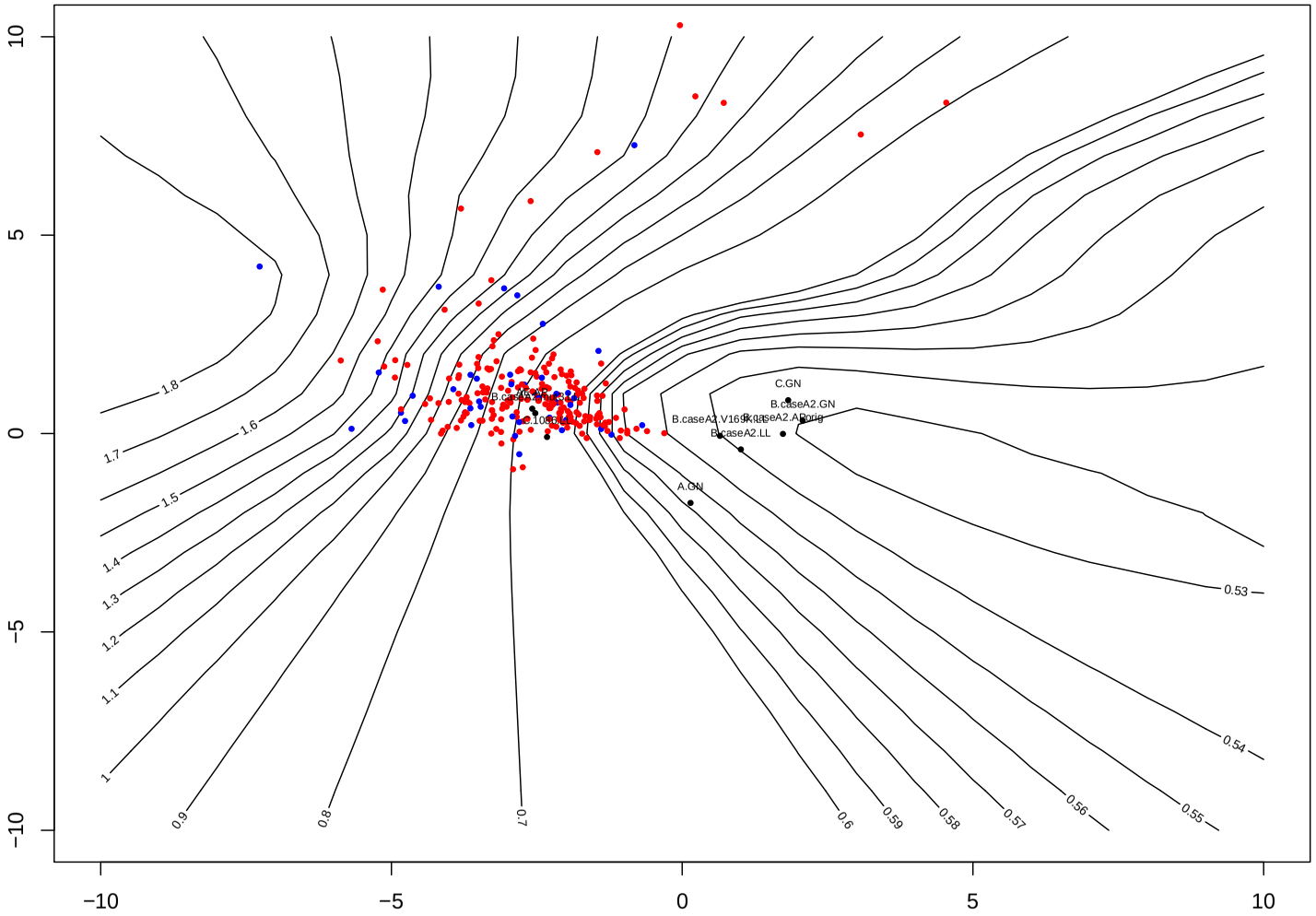

Figure 2.4: Antigenic map of nine V1V2-scaffolded proteins and 246 RV144 case control sera. Blue dots are 41 cases, red dots are 205 controls, and black dots are the nine antigens. The contour lines display the odds ratio from a correlates of risk analysis for each point on the map (see Section 2.3.2 for details).

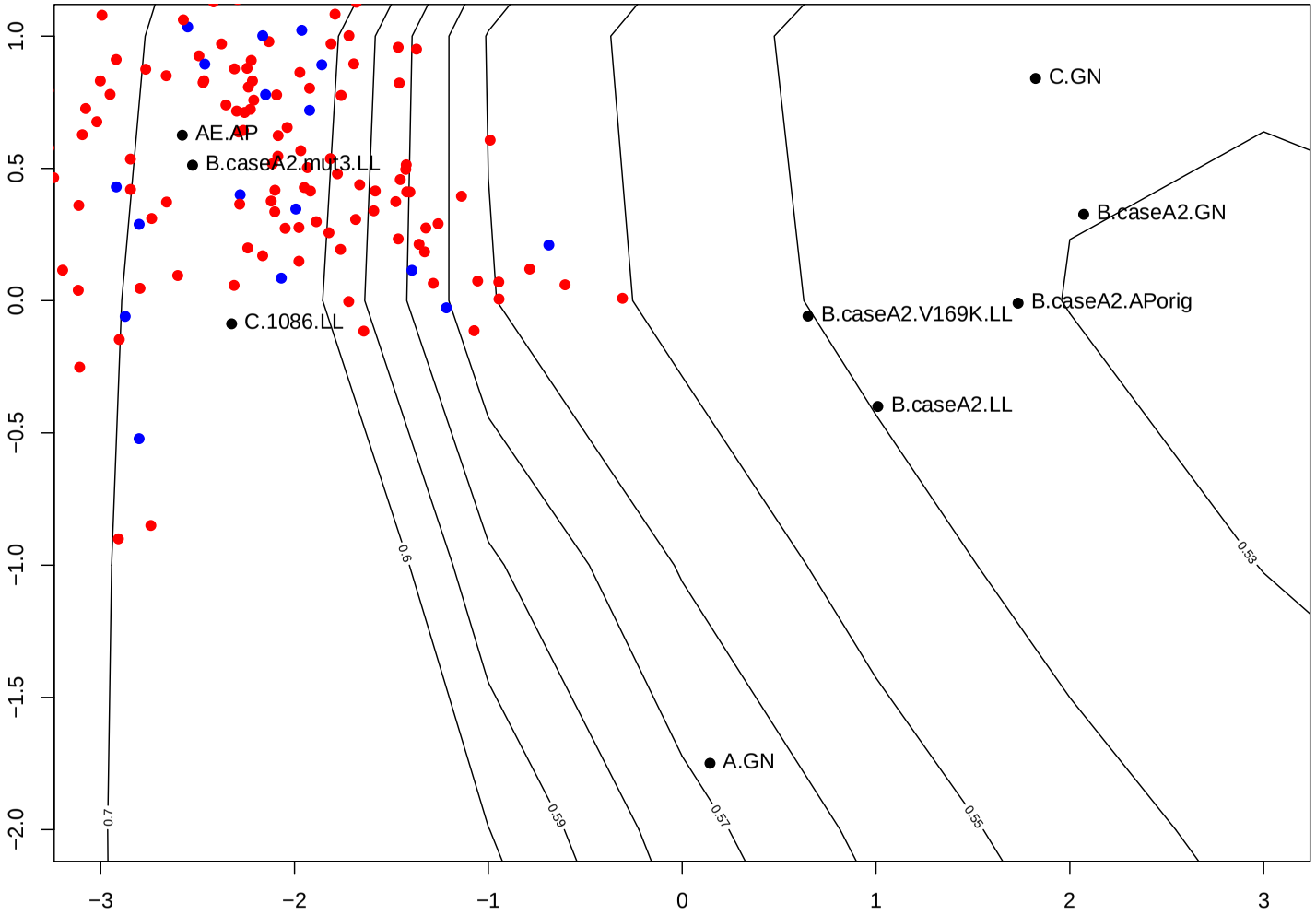

Figure 2.5: Close up view of antigenic map of nine V1V2-scaffolded proteins and 246 RV144 case control sera showing all nine protein locations. Blue dots are 41 cases, red dots are 205 controls, and black dots are the nine antigens. The contour lines display the odds ratio from a correlates of risk analysis for each point on the map (see Section 2.3.2 for details).

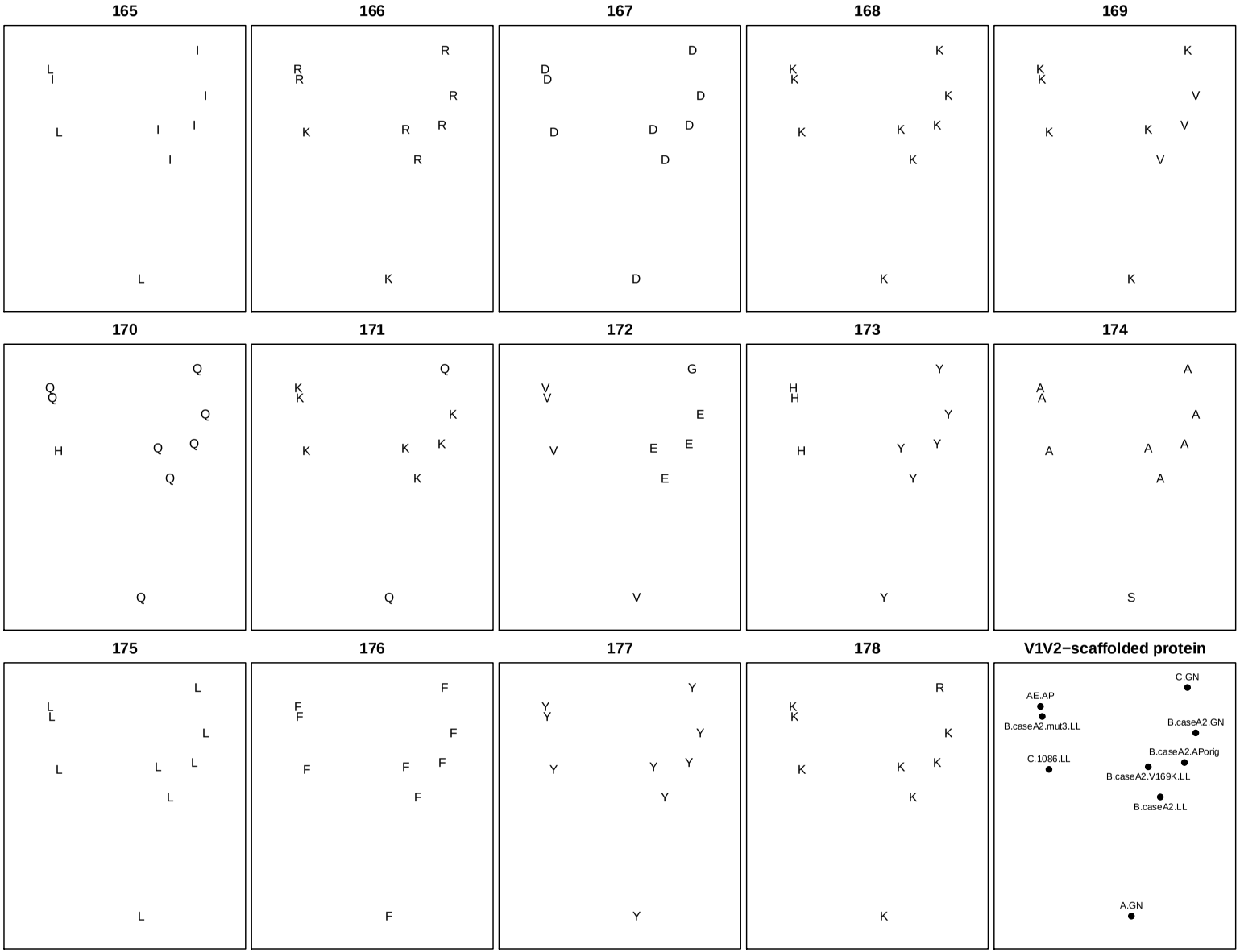

Figure 2.6: Amino acid distribution for V2 hotspot positions HXB2 165 to 178. Each panel shows the 1-letter amino acid abbreviation at a given position for each of the nine V1V2-scaffolded proteins. The location in the panel determines the protein for each amino acid. For reference, the panel in the lower right corner labels the position of each protein within the map.

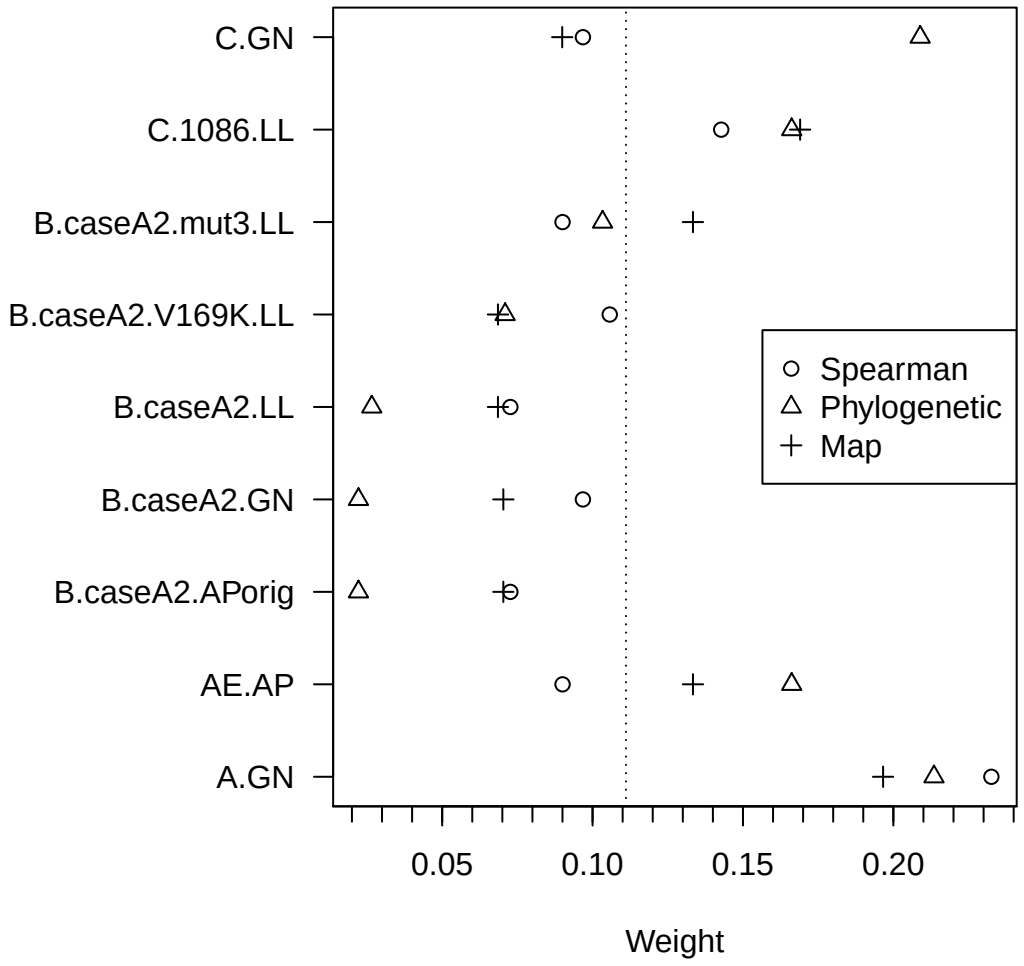

Figure 2.7: Weights for the Spearman correlation, phylogenetic and antigenic map methods for combining the nine V1V2-scaffolded protein readouts into a single biomarker. The dotted vertical line shows the effective weights for the average readout across the nine proteins.

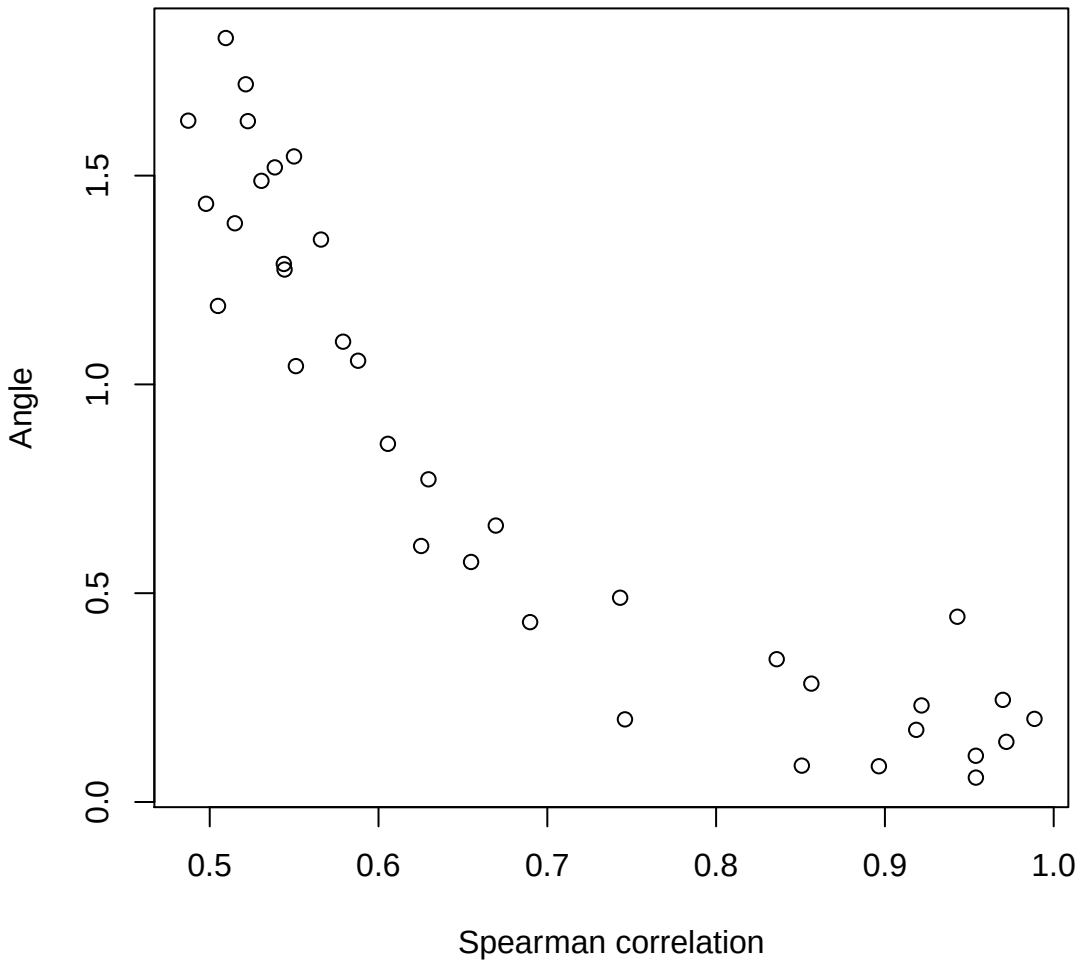

Figure 2.8: For each antigen pair, the relationship between Spearman correlation and the angle (using the medoid of the sera locations as the vertex) within the antigenic map are shown.

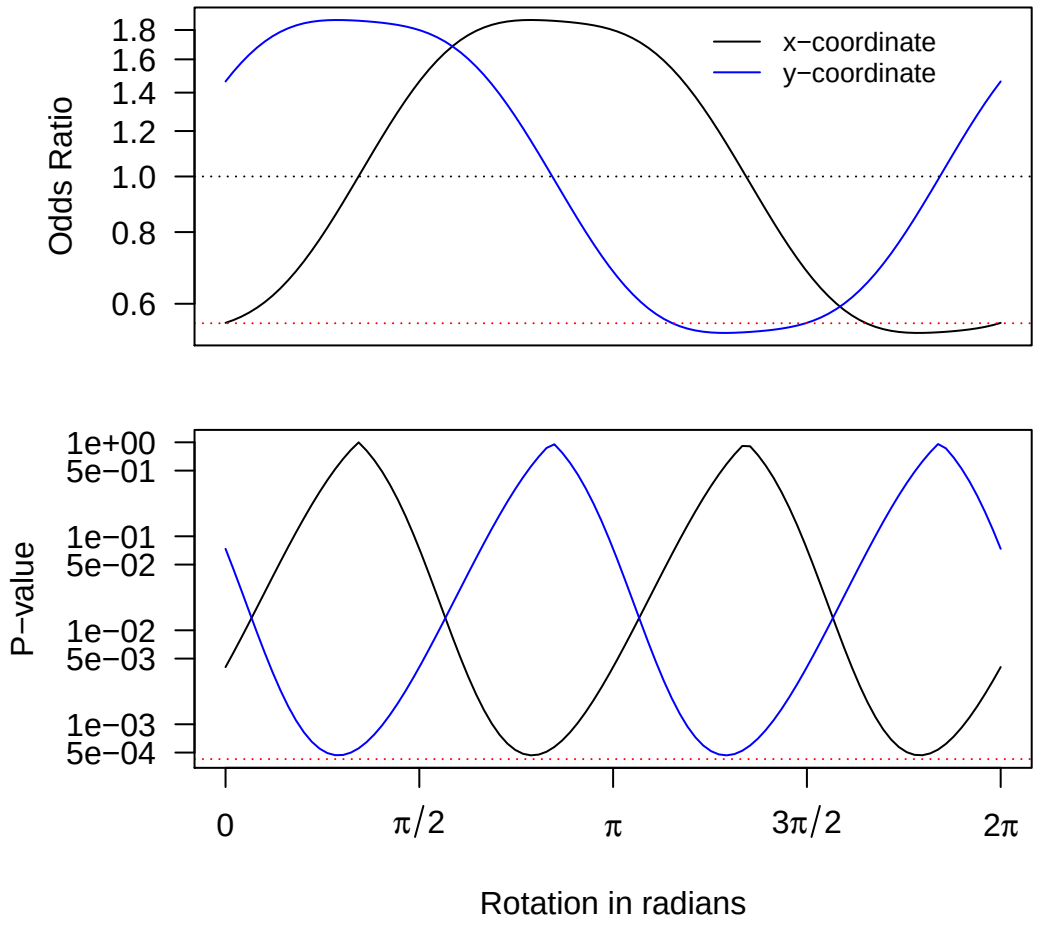

Figure 2.9: The black line in the top panel shows the odds ratio associated with a 1 SD change in the $x-$coordinate (and $y-$coordinate in blue) for each rotation of the antigenic map between 0 and 2π radians. The dotted gray line indicates an odds ratio of 1 while the dotted red line indicates the odds ratio based on the average readout across all nine antigens. The black line in the lower panel shows the p-value for the odds ratio for the $x-$coordinate (and $y-$coordinate in blue) for each rotation of the antigenic map between 0 and 2π radians. The red dotted line shows the p-value for the odds ratio for the average readout biomarker.

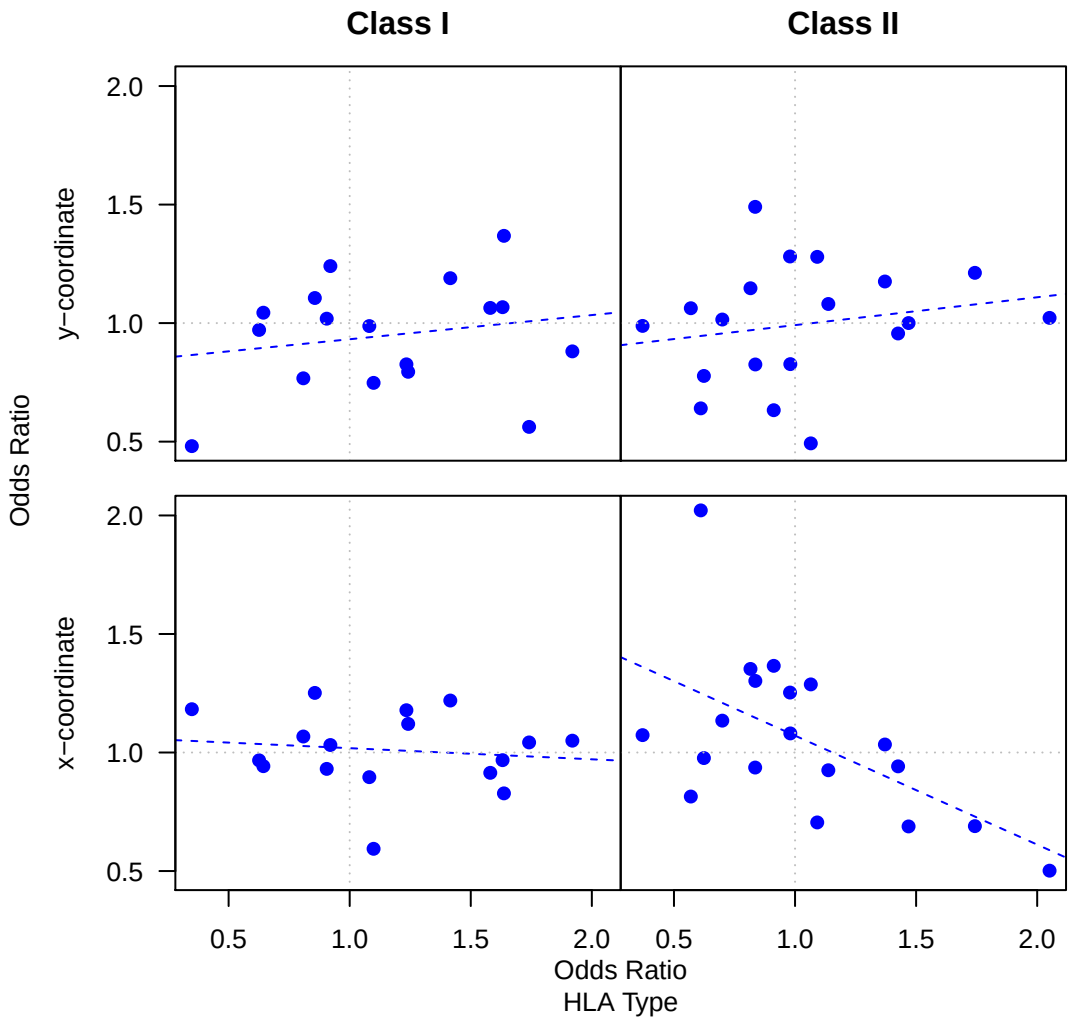

Figure 2.10: The top two panels show a scatter plot of the odds ratios for the association of the $y-$coordinate with infection risk versus the odds ratio for the association of HLA Class I (top left panel) and HLA Class II (top right panel) with infection status. Similarly the bottom two panels show these same plots using the odds ratios for the association of the $x-$coordinate with infection risk.

V1V2-scaffolded protein
A.GN
AE.AP
B.caseA2.APorig
B.caseA2.GN
B.caseA2.LL
B.caseA2.V169K.LL
B.caseA2.mut3.LL
C.1086.LL
C.GN

Table 2.1: Listing of the nine V1V2-scaffolded proteins used to generate the RV144 antigenic map of V2 binding antibodies.

Method	Odds Ratio	P-value
Average	0.555	0.00043
Spearman	0.557	0.00049
Phylogenetic	0.566	0.00047
Map	0.572	0.00066

Table 2.2: Estimated odds ratios per 1 SD change in the combined biomarker along with p-values.

	HLA		Coordinate		
	N	OR	OR_x	OR_y	P-value
Class I					
B:44	28(5)	1.10 (0.79)	0.59 (0.0290)	0.75 (0.37)	0.0695
B:13	39(9)	1.64 (0.25)	0.83 (0.3473)	1.37 (0.05)	0.0778
B:40	41(3)	0.35 (0.11)	1.18 (0.4619)	0.48 (0.09)	0.1174
B:18	33(8)	1.74 (0.21)	1.04 (0.8667)	0.56 (0.18)	0.2460
Cw:01	80(15)	1.24 (0.59)	1.12 (0.4933)	0.79 (0.23)	0.2961
Cw:03	75(11)	0.81 (0.71)	1.07 (0.6961)	0.77 (0.20)	0.3191
B:46	69(13)	1.23 (0.57)	1.18 (0.3488)	0.83 (0.33)	0.3240
Cw:08	48(10)	1.42 (0.39)	1.22 (0.3102)	1.19 (0.29)	0.3938
Cw:04	57(9)	0.92 (1.00)	1.03 (0.8596)	1.24 (0.17)	0.4040
B:51	27(4)	0.86 (1.00)	1.25 (0.3519)	1.11 (0.60)	0.5939
A:24	99(22)	1.92 (0.08)	1.05 (0.7540)	0.88 (0.43)	0.6567
Cw:07	128(22)	1.08 (0.87)	0.90 (0.4647)	0.99 (0.93)	0.7645
B:15	75(16)	1.58 (0.20)	0.91 (0.5833)	1.06 (0.69)	0.7687
A:11	119(24)	1.63 (0.17)	0.97 (0.8267)	1.07 (0.65)	0.8688
A:02	112(15)	0.64 (0.23)	0.94 (0.6915)	1.04 (0.77)	0.8688
A:33	45(7)	0.91 (1.00)	0.93 (0.7127)	1.02 (0.93)	0.9267
B:58	26(3)	0.63 (0.59)	0.97 (0.8871)	0.97 (0.90)	0.9843

Table 2.3: Table of odds ratios for each HLA Class I allele represented by at least 10% of the 246 RV144 case control subjects. Under HLA, N is the number of total subjects with the allele, the number of infected subjects with the allele is given in parentheses. Also under HLA, OR is the odds ratio for an association of the allele with infection status based on the contingency table determined by N. The Coordinate columns show the odds ratio OR_x (and OR_y) for an association between the $x-$coordinate (and $y-$coordinate) and an indicator of the allele. In parentheses next to each odds ratio is the p-value testing for a departure from an odds ratio of one. The P-value column is an overall p-value testing both coordinates simultaneously. The alleles are sorted by P-value.

	HLA		Coordinate			
	N	OR	OR_x	OR_y	P-value	Adj.
Class II						
DPB1:13	88(11)	0.61 (0.22)	2.02 (0.0003)	0.64 (0.10)	0.0001	0.001
DRB1:07	42(11)	2.05 (0.11)	0.50 (0.0014)	1.02 (0.92)	0.0041	0.073
DPB1:04	80(14)	1.09 (0.86)	0.71 (0.0317)	1.28 (0.11)	0.0146	
DRB1:12	76(12)	0.91 (0.86)	1.37 (0.0851)	0.63 (0.08)	0.0165	
DPA1:01	148(29)	1.74 (0.16)	0.69 (0.0248)	1.21 (0.29)	0.0286	
DRB1:09	46(8)	1.07 (0.83)	1.29 (0.2641)	0.49 (0.09)	0.0641	
DQB1:05	147(23)	0.84 (0.61)	0.94 (0.6827)	1.49 (0.07)	0.0859	
DQB1:02	57(12)	1.47 (0.32)	0.69 (0.0372)	1.00 (1.00)	0.1078	
DPA1:02	198(32)	0.84 (0.67)	1.30 (0.1482)	0.83 (0.22)	0.1231	
DRB1:15	91(15)	0.98 (1.00)	1.25 (0.1582)	1.28 (0.11)	0.1237	
DPB1:05	87(11)	0.62 (0.28)	0.98 (0.8812)	0.78 (0.14)	0.2876	
DQB1:03	139(23)	0.98 (1.00)	1.08 (0.6084)	0.83 (0.21)	0.3430	
DPA1:04	28(4)	0.82 (1.00)	1.35 (0.2196)	1.15 (0.49)	0.3931	
DRB1:11	28(3)	0.57 (0.59)	0.81 (0.3567)	1.06 (0.78)	0.6060	
DQB1:06	49(10)	1.37 (0.52)	1.03 (0.8592)	1.18 (0.31)	0.6211	
DRB1:14	44(8)	1.14 (0.82)	0.93 (0.6843)	1.08 (0.65)	0.8075	
DRB1:04	39(5)	0.70 (0.64)	1.13 (0.5315)	1.02 (0.93)	0.8192	
DPB1:02	74(15)	1.43 (0.35)	0.94 (0.7138)	0.96 (0.78)	0.9100	
DRB1:03	27(2)	0.37 (0.27)	1.07 (0.7586)	0.99 (0.96)	0.9493	

Table 2.4: Table of odds ratios for each HLA Class II allele represented by at least 10% of the 246 RV144 case control subjects. Under HLA, N is the number of total subjects with the allele, the number of infected subjects with the allele is given in parentheses. Also under HLA, OR is the odds ratio for an association of the allele with infection status based on the contingency table determined by N. The Coordinate columns show the odds ratio OR_x (and OR_y) for an association between the $x-$coordinate (and $y-$coordinate) and an indicator of the allele. In parentheses next to each odds ratio is the p-value testing for a departure from an odds ratio of one. The P-value column is an overall p-value testing both coordinates simultaneously. The alleles are sorted by P-value. The last column shows Holm adjusted p-values (adjusting the overall p-values) for the two alleles that have an adjusted p-value of less than 0.10.

Chapter 3

PANEL SELECTION

3.1 Introduction

In many immunology assays the platform allows an investigator to run samples against a large panel of antigens. For reasons of cost, time, and limited sample availability the investigator is typically constrained by the number of antigens that can be run. For this reason we are interested in down selecting from a large panel of potential reagents to a smaller panel. When selecting this subset of antigens we would like to capture as much of the information as possible, using the smaller panel, that we would have collected had the full set of antigens been analyzed.

As a result of the RV144 immune correlates study it has been hypothesized that IgG antibodies directed to the V1V2 loop may be a correlate of protection. At least some of these antibodies are directed towards $\beta-$strand C [50] of the V2 loop; a region proximal to the $\alpha 4\beta 7$ binding motif which has been implicated in HIV-1's ability to infect activated CD4+ T-cells in the mucosa [53, 1].

The V1V2 antigen used in the RV144 correlates of risk analysis is a scaffolded construct that is thought to be properly glycosylated and maintain the folded structure of the V1V2 loop. This reagent consists of the V1V2 peptide sequence of the HIV-1 strain Case-A2 scaffolded to a Murine leukemia virus (MuLV) envelope protein gp70 [62, 61]. This is in contrast to many other V2 antigens that were analyzed as secondary variables in the RV144 correlates analysis which were either denatured linear peptides or cyclic peptides. Another important feature of the primary V1V2 antigen is that it is based on a subtype B sequence. While the vaccine used in RV144 contained components from both subtype B and subtype CRF01_AE strains, infections occurring during the RV144 trail were predominantly subtype CRF01_AE which accounted for 89% of the infections. Since the assay readouts from this antigen were identified as a correlate of risk, there has been an interest in creating new V1V2-scaffolded constructs that present the V1V2 loop from additional strains to better understand the properties of the RV144 vaccine-elicited V1V2 binding antibody response. Not only will these antigens be run in additional assays of the RV144 case-control samples but they will be used in future trials as a potential benchmark in HIV-1 vaccine development. To select among the newly generated V1V2-scaffolded proteins, a down selection process

was undertaken in order to generate a small number of antigens that could be run on the precious RV144 case-control samples with the goal of providing as much additional information regarding the V1V2 response as possible. This down selection process was also used to inform the choices regarding antigen reagents for trials in the planning stages.

Another example of a panel selection problem is the Neutralization Serotype Discovery Project (NSDP). In many licensed vaccines the mechanism of protection is either thought or know to be neutralizing antibodies [64]. A primary goal of HIV-1 vaccine design is to develop a vaccine that can elicit broadly neutralizing antibodies. Natural infection is one means to study these antibodies. Although antibodies generated through natural infection typically only strongly neutralized autologous virus that is older than the sera that contains the neutralizing specificity [68], monoclonal antibodies (mAbs) have been identified that are considered broadly neutralizing. A major barrier to designing an effective HIV-1 vaccine is to overcome the strain specific neutralization response often seen in natural infection and generate an immune response which generates neutralizing antibodies to a wide range of globally circulating strains. In the NSDP a matrix of neutralization assay readouts were generated using approximately 200 sera and 200 viral isolates. One goal of this project is to identify one or more sub panels that can efficiently predict the neutralization of future sera (i.e., sera obtained from a vaccinated trial participant or from an individual infected with HIV-1 through natural infection) to currently circulating viruses.

3.1.1 Antigenic Space

In any panel selection problem there is a need to identify the antigenic space of interest. This antigenic space may be specific to an antigenic region such as in the shape space [60] defined by physiochemical properties of the interaction between antibodies which target a specific epitopic region (e.g., $\beta-$strand C of the V2 loop). Alternatively, it may be the entire neutralization space of HIV-1 which is know to encompass multiple antigenic specificities such as the CD4 binding site, CD4 induced region, V3, specific glycosylation patterns, the membrane proximal external region (MPER) of gp41, the PG9/16 region and potentially dozens of others. The antigenic space is in part determined by immunological function and

need not be restricted to antibody specificities. Other examples of antigenic spaces include CD4 or CD8 specific antigenic targets. The scientific aim of interest determines which immunological assay and which reagents will be used. The assay and reagents will necessarily be selected from existing technologies and existing reagents although what is available may not be sufficient to perfectly capture the scientific aim of interest. For example, one aim of interest for the NSDP is to assess vaccine elicited, transmission relevant, neutralization magnitude and breadth against worldwide circulating strains. Our ability to address this scientific aim is likely to be limited by the available target viruses, the use of neutralizing antibodies generated in response to natural infection, and the fact that neutralization is measured in vitro rather than at the point of infection. Nevertheless, we use the available data to assess our aims. Additionally, within the antigenic space of interest we are often especially interested in a subspace. For example, in the NSDP the space of interest is neutralization of circulating strains worldwide whereas we are also particularly interested in the subspace of circulating strains specific to the upcoming efficacy trials in the Republic of South Africa (RSA).

There are three categories of reagents used within an assay; 1) antigenic reagents or antigens, 2) specimen samples, and 3) assay specific reagents. In this chapter we are interested in the first category and down selecting the set of antigens that will be most informative across the second category, the target specimens. The third category of reagents (e.g., those used to generate standard curves or the capture antibody used within an ELISA assay) are not discussed further and for our purposes are considered a part of the assay operating procedures.

Once we identify the antigenic space of interest, the panel selection problem reduces to a comparison of the coverage provided by a subset of K antigens versus the coverage provided by the entire set. With larger K presumably the coverage is better and the trade off is simply coverage versus the cost to run an experiment using K antigens.

In this chapter we present panel selection criteria and methods applied to the two case studies mentioned above. In Section 3.2 we present the V1V2-scaffolded protein panel selection problem and in Section 3.3 we present the NSDP panel selection problem.

3.2 V1V2-scaffold Protein Panel Selection

As discussed briefly above, the V1V2 scaffold panel selection problem is to identify a panel of V1V2-scaffolded antigens that will be tested as correlates of risk in the RV144 study and as a benchmark for future HIV-1 vaccine trials. Although the V1V2 region of the HIV-1 envelope protein is fairly large (typically about 100 amino acids in length) and has the potential for multiple antigenic targets, there is evidence that the RV144 vaccine induced a binding antibody response that was focused on a linear portion of $\beta-$strand C of the V2 loop [12, 89] which is roughly the size of an antibody footprint. This suggests that the antigenic space of interest is a single antigenic target which can be thought of in terms of shape space as proposed by Perelson and Oster in 1979 [60] and presented in detail in Chapter 2. In this section we detail the criteria and methods used to select the antigens that were used in that analysis.

This panel selection exercise consists of down selecting from a list of 20 scaffolded antigens run in an ELISA assay against plasma samples collected at weeks 0 and 26 from 32 vaccine and 8 placebo recipients in the RV144 trial. Four of these antigens used a scaffold know as J08, nine used a Histidine tag scaffold, and the remaining seven used the gp70 envelope protein from MuLV. In addition, two controls were run, one without antigen added to the well and the other consisting of only the MuLV scaffold protein gp70. Interestingly, J08 scaffolded to an SIV V1V2 loop region was intended as a control for the J08-scaffolded antigens but there were clear vaccine induced responses to this antigen demonstrating the high level of cross-reactivity of the antibodies generated by the RV144 vaccine to HIV-1 and related viruses. The ELISA assay was run at six 3-fold serial dilutions between 1:100 and 1:24300.

There are several criteria that can be used to select a panel of K antigens out of the 20 that were are available. We consider two basic classes of requirements for antigen selection: 1) antigen assay characteristics and 2) position of antigen in the antigenic space of interest. The reason for the first class of requirements are purely technical. If we did not have any uncertainty in the assay we would make a decision based solely on the second criteria.

3.2.1 Antigen assay characteristics

Various antigens may have different operational characteristics within an assay. We first need the antigen to be able to detect a vaccine effect. For this criterion we can use the p-value from a statistical test comparing peak immunogenicity by treatment arm to compare antigens. We also want to have a high signal-to-noise ratio, defined as the standard deviation of the subject specific readouts divided by the mean of the standard deviation of the within subject replicates. The inter-quartile range (IQR) is also of interest as a measure of the dynamic range of the assay. We are interested in a large dynamic range of the assay readout although purely selecting on this criteria may be misleading as a readout with large dynamic range may also have a low signal-to-noise ratio. Similarly we are interested in antigens with a high positive response rate. Response rate alone may not be that helpful as we may not have a well developed positivity criteria or we may have difficulty distinguishing between vaccine induced response and existing immunity. In practice we consider the positive response rate in conjunction with P-values from a comparison of readouts by treatment arm. As an example, for each dilution, the signal-to-noise ratio and p-values from a Wilcoxon rank sum test comparing the week 26 readouts by treatment assignment are plotted in Figure 3.1.

3.2.2 Position of antigen within antigenic space

One way of comparing two antigens within antigenic space is to compare their correlations across readouts. In Figure 3.2, we look at the matrix of 22×22 pairwise Spearman correlations between the 20 scaffold antigens plus the 2 negative controls across the 32 vaccine recipient baseline adjusted 1:100 readouts at week 26.

In an ideal framework the shape space of the C-beta strand directed antibodies induced by RV144 would be recovered using multidimensional unfolding (MDU) using the 20×32 matrix of week 26 vaccine recipient ELISA readouts. In this space we get the coordinates of both the antigens and the antibodies and we could in theory pick the antigens that best detect the variation in location of the antibodies. The critical information that is missing is that we do not know which component of variation, either a single dimension or potentially some combination of dimensions, might confer protection. We might infer the components

of variation that determine protection based on the location of antigens that closely match circulating strains. Figure 3.3 is a hypothetical 2-dimensional antigenic map of the shape space defined by V2 antigenic region of interest. The blue circle encompasses the vaccine induced antibody responses and the orange region encompasses circulating strains. For each antibody on the map we assume that the protective effect of that antibody against an exposing strain is a decreasing function of the distance on the map between the antibody and the strain. In the simplest case, all strains within a fixed distance would be completely protected against while beyond that distance there would be no protection. In this hypothetical map the primary RV144 IgG V1V2 correlate is positioned in the ideal location to detect antibodies with the most overlap with circulating strains. In reality, this is unlikely to be the case and additional antigen probes may provide additional information about the overlap between vaccine induced antibodies and circulating strains. The probe in the lower left corner is orthogonal to the primary antigen in the sense that it would provide no additional information about which vaccine recipients are more likely to be protected. Additionally, if the antigenic map has more than 2 dimensions there may not be one ideal direction with which to measure the protective effect of antibodies and additional probes may be required to completely characterize a protective effect of vaccination.

This provides us with a second way of comparing two antigens within antigenic space which is to look at the angle in the antigenic map between the two antigens using the centroid of the antibodies from the vaccine group as a vertex. Figure 3.4 shows a heatmap of the pairwise angles based on a six dimensional antigenic map of the baseline adjusted 1:100 readouts from V1V2-scaffolded antigens.

3.2.3 Panel selection based on clusters

Now that we have identified assay based and antigenic space based criteria for selecting a panel a reasonable approach would be to cluster the antigens into K groups based on one of the antigenic space based criteria then select one antigen from each cluster based on a combination of all of the criteria specified above. In this way we would form a panel of K antigens. The assay based criteria can be used either as a screen prior to clustering or as

a means of breaking ties between antigens within the same cluster that would both appear to provide similar information in antigenic space if selected.

Table 3.1 shows the results of 10 clusters based on baseline adjusted 1:100 dilution. Hierarchical clustering trees using complete linkage based on a dissimilarity measure defined as one minus the Spearman rank correlation coefficient between Week 26 readouts of vaccine recipients were used to cluster antigens into common antigenic groups. Dendrograms displayed on the heatmaps in Figure 3.2 visually display the clustering. No antigen and MuLV are included in the table as clusters of size one but are not of primary interest. The columns shown in the table are:

- Cluster: a numeric id for the cluster that separates each group of antigens
- Correlation with p565f: the Spearman correlation with the V1V2 antigen used in the RV144 correlates analysis
- Signal-to-noise ratio: the standard deviation of the subject specific readouts divided by the mean of the standard deviation of the within subject replicates
- Response Rate: based on a threshold defined as 3 standard deviations above the mean of the week 0 readouts
- IQR: difference between upper and lower quartiles followed by inter-quartile range in brackets
- P-value: P-value from a Wilcoxon rank sum test comparing readouts by treatment group
- Angle: Angle between p565f and the specified antigen with vertex at the mediod of the plasma in an antigenic map of the V1V2 scaffold antigens and 32 vaccine plasma samples.

3.2.4 Conclusions V1V2 panel selection

Ultimately the antigens that were used for follow-up analyses of the RV144 case-control samples were selected based on the tables and figures described above using a combination of the criteria summarized in Table 3.1 along with a variety of competing factors. One competing factor was the interest in replicating the original result and the final list of antigens that were run on case-control samples included three variants of the original Case A2 antigen produced in three different labs. Another competing factor was the interest in further understanding the results of the RV144 V2 specific sieve analysis [70] and experiments conducted on V2 specific monoclonal antibodies elicited by the RV144 vaccine sera [46] which resulted in the addition of two mutated variants of the original Case A2 strain antigen. Additionally, interest in upcoming efficacy trials in RSA influenced the choice of antigens. These additional criteria, or similar external factors, are likely to play a role in any antigen selection process.

At the same time there is a case to be made for selecting antigens based purely on the selection criteria outlined in the previous section. The antigenic map described in Chapter 2 shows that many of the antigens selected probe antigenic space from the same general angle (Figure 2.4). This may in fact be a feature of the antigenic space but it is also possible that there are antigens that would generate a more accurate map if they had been included. One hypothesis is that an antigenic map would be best captured if the antigens within the space are placed orthogonally to one another with respect to the medoid of the sera responses. In the six dimensional antigenic space we generated for the down selection process three J08 antigens based on A, B, and C subtype strains are nearly orthogonal to the original Case A2 antigen (Table 3.1). One reason these were not selected was due to the poor response rate of these antigens but despite the poor response rates they may have be good choices for generating an antigenic map using the case-control sera. Similarly, the SIV antigen was not chosen. Again, there were additional factors when selecting the antigens beyond the goal of capturing the antigenic space of the RV144 elicited V1V2 response. However, the SIV antigen clearly detects vaccine elicited responses and occupies a unique position in antigenic space shown both in the heatmap based on Spearman correlations and the heatmap based

on angles within the antigenic map (Figures 3.2 and 3.4).

3.3 Neutralization Serotype Discovery Project Panel Selection

Our goal is to identify a small panel, between three and twelve viruses, that predict neutralization to the larger population of NSDP isolates. This panel will be useful for developing immunogenicity endpoints that capture magnitude and breadth (MB) of neutralization to circulating strains in HIV-1 clinical trials. Mascola et al. in 2005 [49] recommend, "the establishment of standardized panels of Env-psuedotyped viruses to assess the potencies and breadths of NAbs elicited by vaccine immunogens," and define a three tier system for assessing the neutralizing antibody response to HIV-1 vaccines. Huang et al. [33] develop statistical methods for the evaluation of the magnitude and breadth of vaccine induced neutralization. Gilbert et al. [21] assess the magnitude and breadth of neutralization to eight tier 1, twenty tier 2, and twenty-seven viruses derived from infected participants in the Vax004 trial. In the RV144 [30] magnitude and breadth of neutralization to a panel of six isolates was chosen as one of six primary immune responses studied in the immune correlates of risk analysis. Seaman et al. [75] analyzed patterns of neutralization sensitivity of 109 molecularly cloned HIV-1 Env pseudoviruses to seven plasma pools. We extend this work to systematically select a small panel of isolates that best predict the overall neutralization of the complete panel of NSDP isolates. We develop a series of isolate selection methods based on clustering and model selection algorithms and compare their performance using a goodness-of-fit criteria. Specifically we apply clustering algorithms to tree based distances and pairwise correlations of neutralization between isolates, we also use the lasso model selection procedure to identify the most predictive isolates. Additionally, we can apply these methods to identify panels of viruses that will predict MB of subtype specific populations of viruses.

3.3.1 Magnitude breadth curves

For each subject we compute a magnitude breadth curve $y = g_i(x)$ for subject i [33]. Given a neutralization magnitude x, y is the fraction of isolates with a neutralization magnitude greater than x. Neutralization magnitude can be measured in many ways. Typically it

is measured as either ID50 or ID80, the dilution at which 50 or 80 percent of viruses are neutralized in the TZM-bl assay. We do not however want to limit ourselves to these readouts and allow any measure of neutralization to be used for the independent variable of the MB-curve. For example, we might want to use the positive area under the dilution curve (pAUC) which may be more sensitive to very low levels of neutralization and removes the added complication of censored values common to ID50 and ID80 readouts. Yu et al. [86] exam both AUC and partial AUC and suggest that partial AUC has advantages of greater power and less complication due to censoring than ID50 and ID80. Although pAUC and partial AUC as defined by Yu are different, they are likely to share similar characteristics particularly with respect to sensitivity to low levels of neutralization.

3.3.2 Statistical methods for panel selection

There are two statistical methodologies that we pursue here; 1) model selection and 2) clustering. We start with a $N \times M$ matrix of neutralization readouts (e.g. pAUC, $\log_{10}$ ID50, $\log_{10}$ ID80, etc.) $\mathbf{X} = (X_{ij})$ where i and j index plasma samples and viral isolates respectively.

Define the weighted mean neutralization score as

$$Z_i = \frac{1}{M} \sum_{j=1}^{M} w_j \cdot X_{ij} \tag{3.1}$$

where w_j are weights. The weights could be used to correct for biases within the set sampled viral isolates but here we use $w_j = \mathbf{1}$ hence Z_i is the area under $g_i(x)$ the magnitude breadth curve for subject i (AUC-MB).

Later in this section we consider neutralization scores defined by sample quantiles as defined in Hyndman and Fan [35]. An estimate of the distribution quantile p for a plasma sample is given by the weighted average of the j^{th} and $(j+1)^{st}$ order statistic of the neutralization readouts for plasma sample i. Define

$$Q_i(p) = (1-\gamma)X_{i(j)} + \gamma X_{i(j+1)} \tag{3.2}$$

where $X_{i(j)}$ is the j^{th} order statistics for plasma sample i, $(j+p-1)/M \leq p < (j+p)/M$, and $\gamma = (M-1)p + j + 1$. We will specifically be interested in the 3 quartiles $Q_i(0.25)$,

$Q_i(0.5)$, and $Q_i(0.75)$ and simplify the notation as $Q1_i$, $Q2_i$, and $Q3_i$. Note that since the MB curve $g_i(x)$ is one minus the empirical distribution function of the neutralization readouts, we have $g_i(Q1_i) = 0.75$, $g_i(Q2_i) = 0.5$, and $g_i(Q3_i) = 0.25$.

Panel selection method 1: lasso

Consider the regression problem of predicting the AUC-MB, $\mathbf{Z} = (Z_1, \ldots, Z_N)$, based on the individual neutralization readouts $\mathbf{X}$. For the NSDP data, we have the number of predictors N approximately equal to the number of independent samples M. There is a circularity inherent to our construction since the outcome is the mean of the predictors. However, since we are interested in identifying a small set of predictors compared to the total number of isolates this is a relatively minor issue. We use the lasso model selection procedure [78] to select a panel. Lasso is a shrinkage method that applies an L^1 constraint to the familiar ordinary least squares (OLS) minimization of linear regression. The beta coefficients are given by

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \tfrac{1}{2} \|\mathbf{Z} - \beta\mathbf{X}\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{3.3}$$

where the size of λ determines the amount of shrinkage. We assume that the $\mathbf{X}$ matrix columns have been scaled to have mean zero and variance one so that isolate selection is not driven by the magnitude of the response.

One property of lasso is model selection since large values of λ tend to make many of the β coefficients shrink to zero. For a given panel size K, we select λ such that all but K of the $\hat{\beta}^{lasso}$ coefficients are zero. After selecting K predictors via lasso we refit the model using OLS based on the non-standardized predictors. This two step process is a variant of the Relaxed lasso [51] known as thresholded lasso and in high dimensional data produces equal or lower prediction loss than the regular lasso estimator. In our context, with approximately equal numbers of predictors and outcomes, we have also seen better prediction based on thresholded lasso than regular lasso.

Panels of three and nine isolates based on the lasso method predicting AUC-MB using pAUC neutralization readouts are shown in Tables 3.2 and 3.3.

Panel selection method 2: Clustering

The second method for panel selection is based on clustering isolates in some space (genetic, antigenic, structural, physiochemical, etc.) then picking a representative isolate from each of K clusters. We then fit a linear regression model to predict the neutralization score given the K predictors. We use the k-medoids algorithm described below to define clusters.

A medoid is defined as a member of a cluster whose average dissimilarity to all of the objects in that cluster is minimal. It is the most centrally located point in the cluster based on the given dissimilarity. The algorithm for selecting k-medoids [41] is similar to the k-means algorithm. The difference is that k-medoids minimizes the sum of pairwise dissimilarities whereas k-means minimizes the sum of squared Euclidean distances. Furthermore, k-medoids uses a data point as the center of each cluster while k-means chooses a point in Euclidean space that does not necessarily correspond to a data point.

We explore four pairwise isolate dissimilarity definitions, or isolate spaces. One is genetic, two are antigenic and another is based on targeted antigenic determinants. We compute dissimilarity matrices based on the following dissimilarities between pairs of isolates:

- Genetic: tree based distance
- Antigenic:
  - Euclidean distance between neutralization readouts
  - one minus Spearman correlation between neutralization readouts
- Targeted antigenic determinants: one minus correlation between IC50 readouts from a panel of 4 monoclonal antibodies (IgG1b12, 2G12, 4E10, 2F5) plus soluble CD4 (sCD4).

Panels of three and nine isolates selected using tree distance based clustering method are shown in Tables 3.4 and 3.5.

3.3.3 Comparing methods of panel selection

We have five methods for selecting a panel K isolates, one using the lasso method and 4 variants of the clustering method based on different isolate spaces. To compare these methods we select panels of sizes $K \in [1, 20]$ and compute the R-squared value for the linear model based on the K selected isolates. We also select 1,000 panels at random for each $K \in [1, 20]$ and compute the R-squared value for each random panel.

Results are shown in Figure 3.5. The gray bands represent the R-squared values from the 1,000 random draws of K-isolates for $K \in [1, 20]$. The dark gray band shows the 2.5 to 97.5 percentile of the R-squared values and the light gray bar extends to the extremes observed over the random draws. Prediction of the mean neutralization is quite good with R-squared greater than or equal to 0.8 using only 3 isolates for all methods except for the method based on targeted antigenic determinants. The Tree based method preforms quite well for all $K \in [1, 20]$ and does not use any neutralization data in predicting clusters. The Euclidean based antigenic distance performs reasonably but not as well as antigenic distance based on correlation. Finally the targeted antigenic determinant based method gives mediocre prediction but the space is not particularly rich with data from only 5 antigenic determinants. A richer space might provide much better prediction.

3.3.4 Cross validation results

To estimate prediction error we use cross-validation (CV). Tenfold cross-validation, as recommended by [29], was chosen as trade off between variance and bias of the K-fold cross validation estimator of prediction error as the number of cross validation sets K are varied. We compute the mean squared error (MSE) of the true neutralization score and our predicted neutralization score using tenfold cross-validation. Figure 3.6 suggests that our predicted values are not overly optimistic. We do see that the difference between the MSE using the full model compared to the MSE based on CV is biggest for the lasso model which is not surprising since lasso uses supervised learning while all of the other methods are unsupervised.

3.3.5 Predicting MB curves

It would be ideal to select a panel of isolates that is able to predict, for each plasma sample, the entire MB curve derived from all of the NSDP isolates. There may be information in the shape of the curve not captured by the area under the curve (AUC-MB). This motivates our next approach to panel selection which is to predict the three quartiles of the neutralization readouts ($Q1_i$, $Q2_i$ and $Q3_i$) for a given plasma sample. Given the three predicted quartiles we can then fit a five-parameter logistic equation to these points fixing the upper and lower asymptotes at 1 and 0 respectively. One parameterization of the five-parameter logistic function with fixed asymptotes is given by

$$g(x) = \frac{1}{[1 + \exp(b(x - e))]^f} \tag{3.4}$$

which is symmetric about the inflection point when $f = 1$ and is asymmetric otherwise. An alternate parameterization with interpretations of each parameter can be found in [26].

3.3.6 Panel Selection using grouped lasso

Motivated by our interest in predicting the three quartiles simultaneously we use the grouped lasso method developed by Breheny and Huang [6] to predict the multivariate neutralization score $Z \equiv (Q1, Q2, Q3)$. The grouped lasso method groups beta coefficients and they either all contribute to the model or they are all set to zero. The method is achieved by minimizing the following L^2 objective function

$$\hat{\beta}^{group} = \underset{\beta}{\operatorname{argmin}} \tfrac{1}{2}\|Z - (X)\beta\|_2^2 + \lambda \sum \sqrt{K_j}\|\beta_j\|_2 \tag{3.5}$$

where j indexes the groups and K_j is the size of each group. Notice that this simplifies to the regular lasso when all of the groups are of size one.

To predict the quartiles simultaneously we group the coefficients in threes and use a block diagonal design matrix. We also scale each quartile to have equal variance which allows each quantile to contribute equally.

In addition to predicting the three quartiles simultaneously using the grouped lasso we also predict each quartile separately using the regular lasso. For a panel size of 9, the Venn

diagram of the selected isolates is shown in Figure 3.7 and a table of the results is shown in Table 3.6.

3.3.7 Results of predicting MB curves

We compare the fit of an estimated MB curve and the observed MB curve defined by the entire set of NSDP isolates using the Cramér-von Mises distance, mean squared distance, between the two curves. For each plasma sample we also compute an optimal predicted MB curve by fitting the logistic curve defined in equation 3.4 to the knots of the observed MB curve. Figure 3.8 shows the results based on the four 9 isolate panels in Section 3.3.6 compared to the optimal predicted curve fits. Isolates selected solely on Q1 prediction appear to perform best on overall MB curve prediction.

3.3.8 Conclusion

In the NSDP panel selection problem a practical number of isolates, approximately nine, can be used to predict both AUC-MB and in fact the entire MB-curve for most sera quite well. A panel using these methods can be used to perform standardized assessments of vaccine elicited neutralization responses.

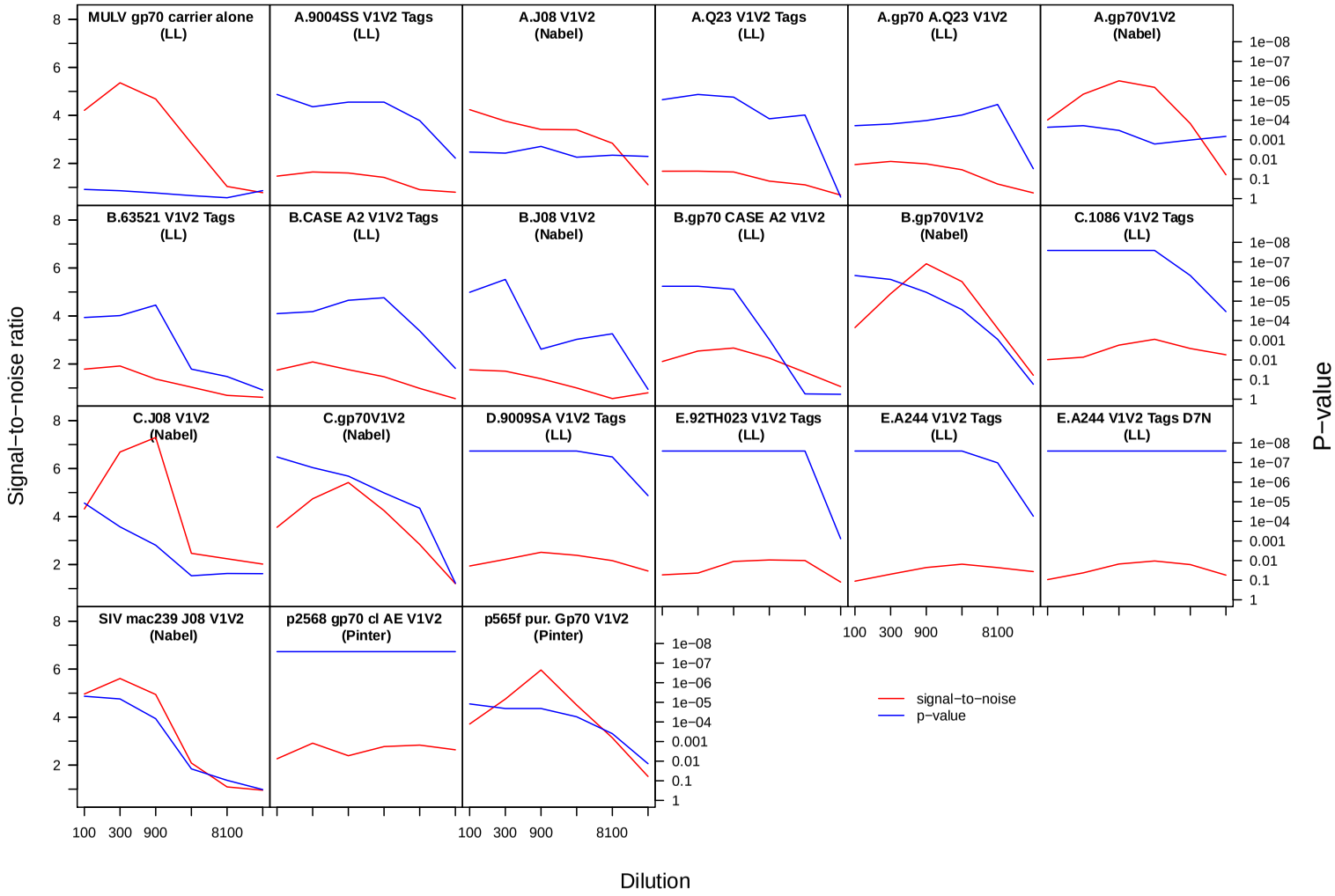

Figure 3.1: The signal-to-noise ratio versus the P-value from a Wilcoxon rank sum test for a treatment effect are plotted in each panel for the 20 V1V2-scaffolded proteins plus the control MuLV gp70 carrier alone. P-values are plotted on the negative $\log_{10}$ scale so small p-values are at the top of the plot and p-values near one are at the bottom of the plot.

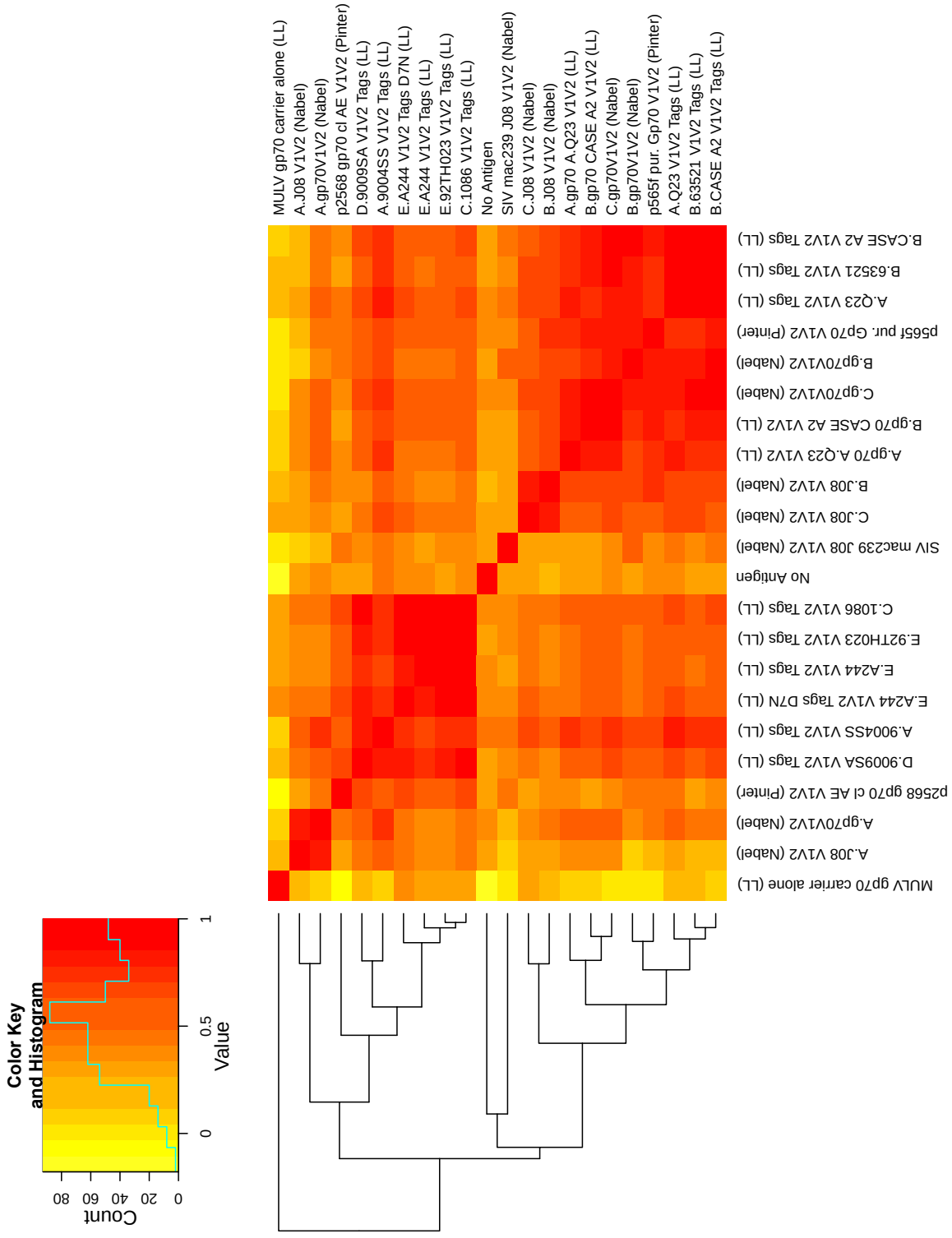

Figure 3.2: A heatmap of Spearman correlations for 20 V1V2-scaffolded proteins plus two controls. The dendrogram on the left shows a hierarchical clustering of the antigens.

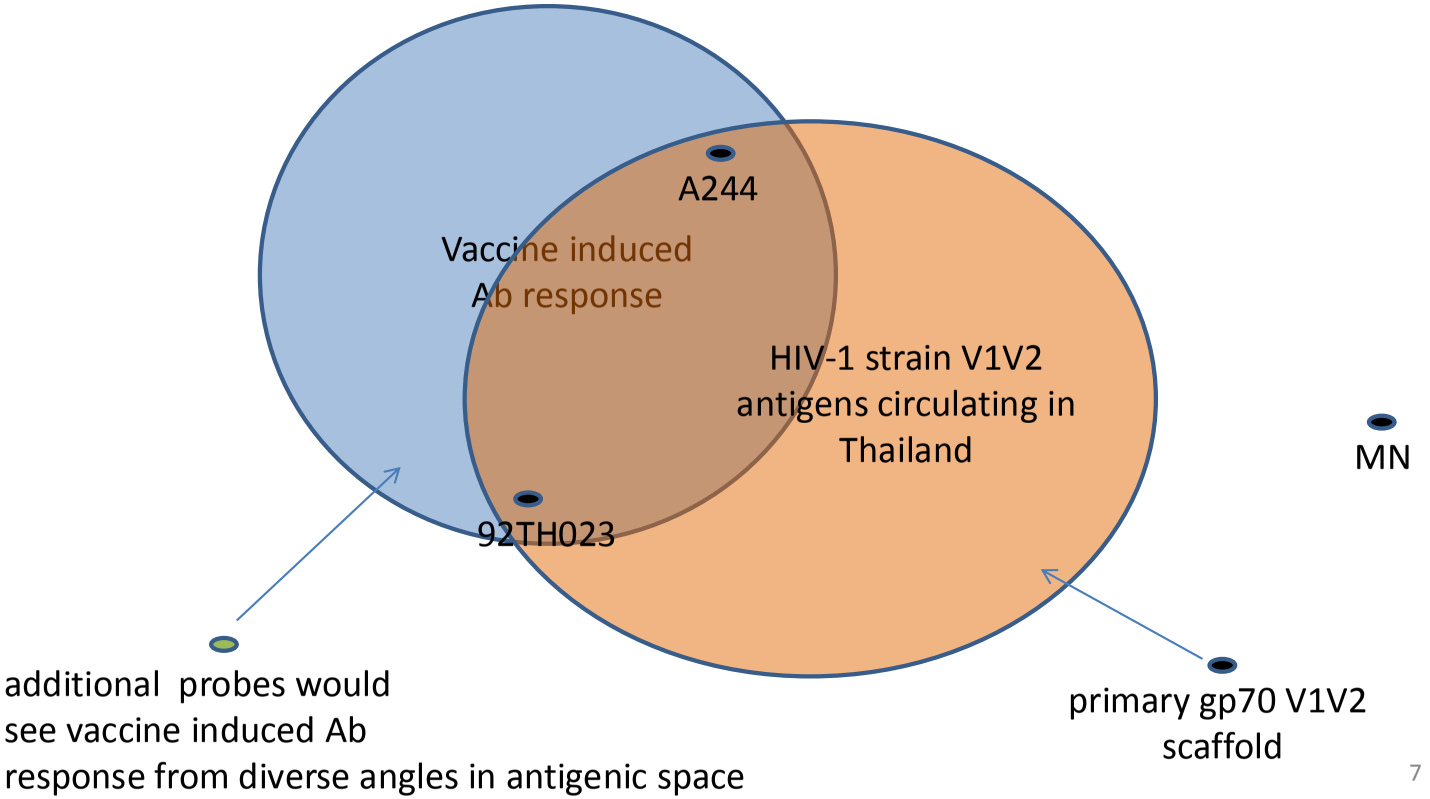

Figure 3.3: Hypothetical antigenic map of the V1V2 antibody responses to the RV144 vaccine. The location of the three RV144 vaccine strains (A244, 92TH023, and MN) along with circulating Thai strains are placed relative to the vaccine elicited responses. In addition, two additional "probes" have been placed on the map; one corresponding to the primary V1V2-scaffolded antigen from the RV144 immune correlates study and another that is orthogonal to that antigen. All locations are hypothetical.

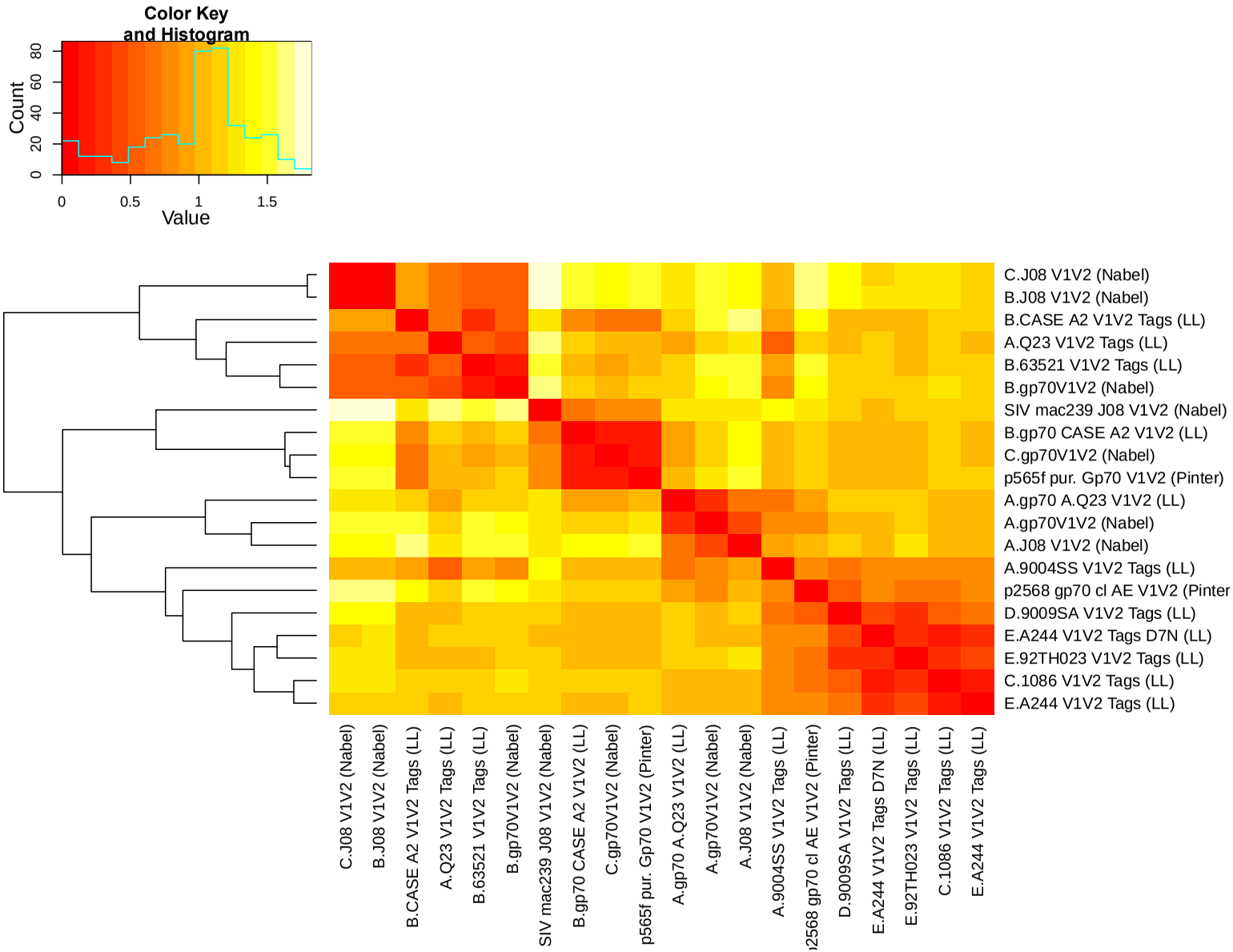

Figure 3.4: A heatmap of the pairwise angles between antigens in the RV144 V1V2-scaffolded down selection study. The vertex of each pair of angles is the medoid of the 32 vaccine group sera locations. The dendrogram shows a hierarchical clustering of the antigens based on the angles. Angles are in radians and a color key at the top left gives the magnitude of each angle.

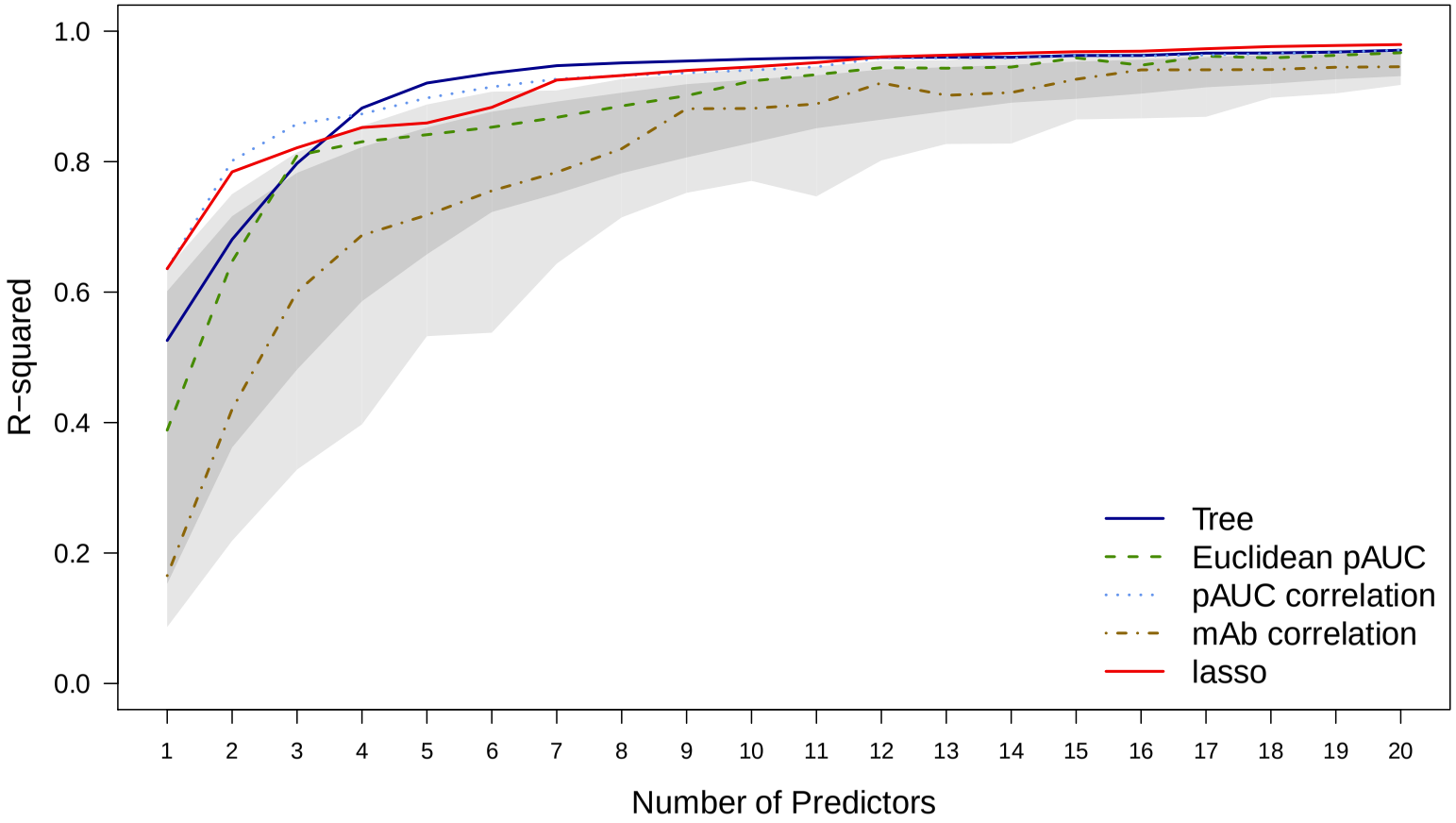

Figure 3.5: Comparison among optimized and randomly selected virus panels. R-squared values for predicting AUC-MB are shown for each of K predictors per panel. Each line corresponds to the optimized K-isolate panels selected using one of the five methods described in Section 3.3.2. The dark gray region shows the range of R-squared values from the middle 95% of random panels. The light gray region shows full distribution of R-squared values across 1,000 random panels.

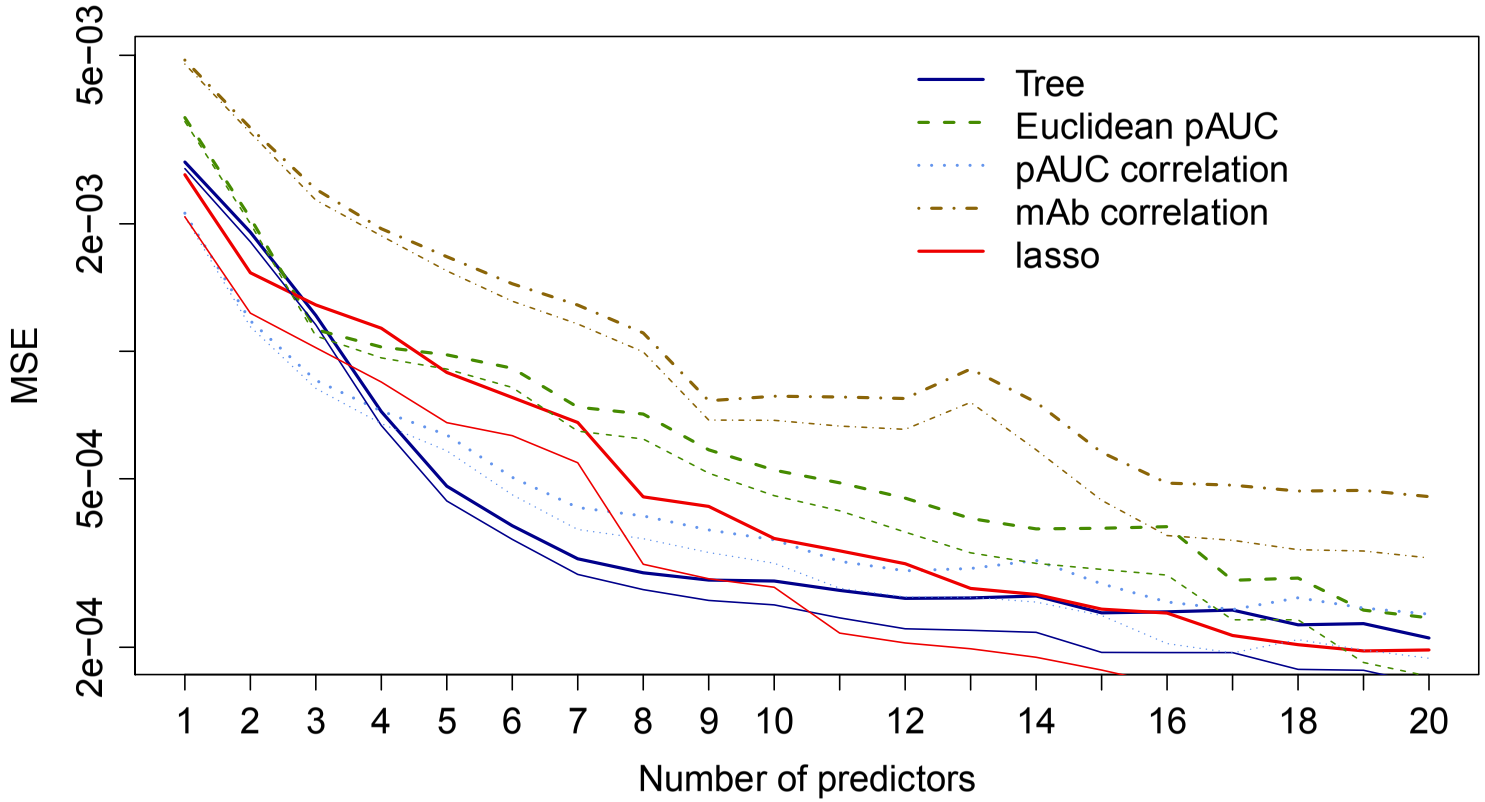

Figure 3.6: Tenfold cross-validation for virus panels of size one to 20. For each method the bold line shows the MSE from a tenfold cross-validation while the narrow line shows the MSE for the original model. The y-axis is log-transformed to better display the differences between methods and the CV and original model results.

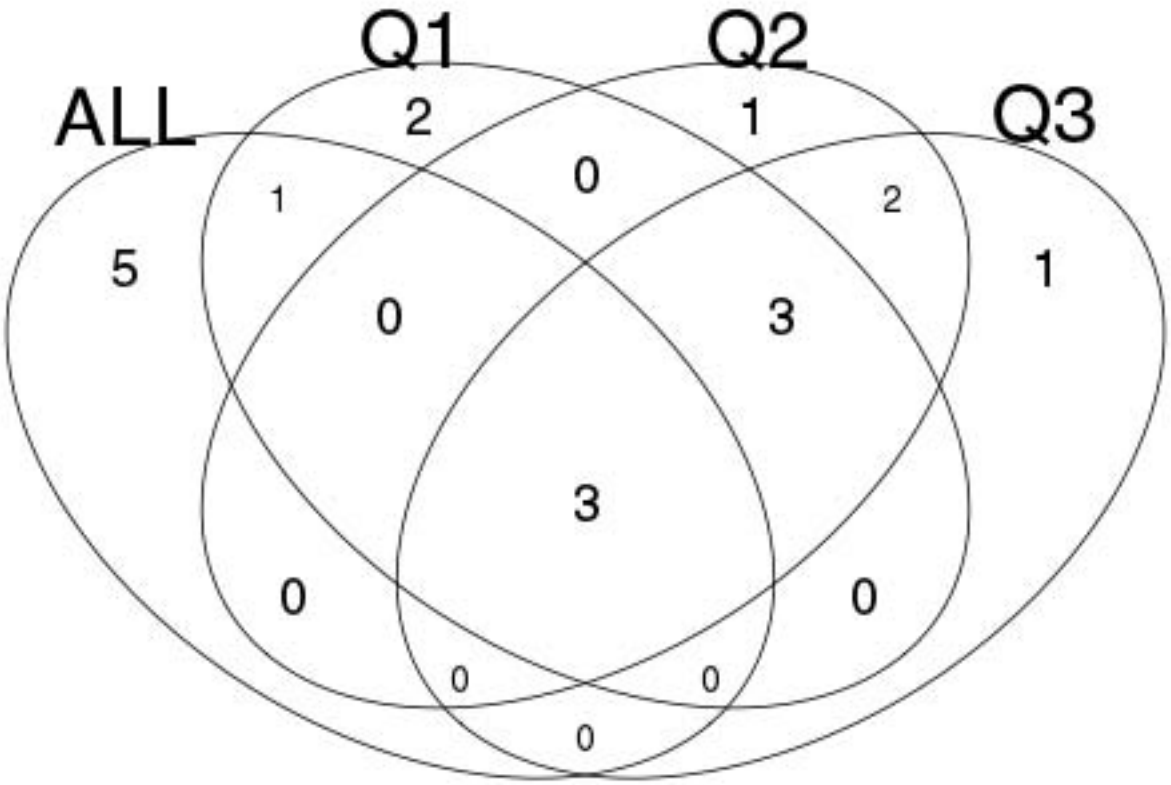

Figure 3.7: Venn diagram comparing 9 isolates chosen using grouped lasso and regular lasso. "All" is the panel chosen by grouped lasso while Q1, Q2, and Q3 are the panels chosen when predicting the individual quartiles using the regular lasso described Section in 3.3.6.

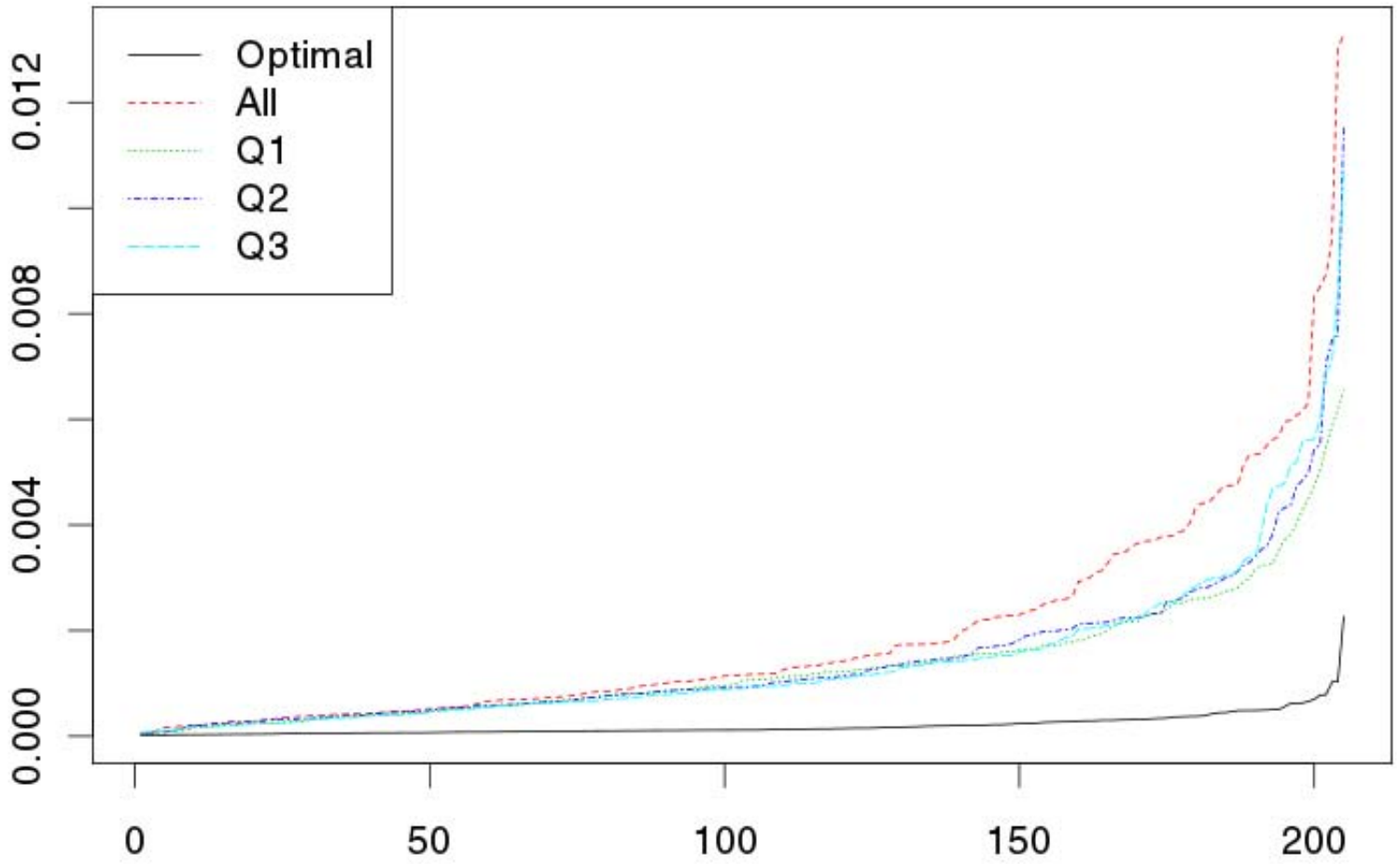

Figure 3.8: The goodness-of-fit between each predicted MB-curve and the observed MB-curve are shown for various predictors for each of the 205 serum samples. The "Optimal" predictor is a logistic curve fitted to all of the knots of the observed MB-curve and shows the best fit we can expect for any logistic curve. The other four lines show the results for the grouped lasso method ("All") and each of the quartile methods described Section in 3.3.6.

Cluster	Correlation	Response	Signal-to-noise	IQR	P-value	Angle
MULV						
MULV gp70 carrier alone (LL)	0.02	3.1%	1.9	0.04 [-0.00,0.04]	0.25	NA
10						
A.J08 V1V2 (Nabel)	0.17	25.0%	4.1	0.48 [0.08,0.56]	0.00056	1.46
A.gp70V1V2 (Nabel)	0.47	31.2%	3.8	0.59 [0.42,1.00]	2.6e-08	1.24
9						
p2568 gp70 cl AE V1V2 (Pinter)	0.50	96.9%	3.0	0.49 [2.37,2.86]	2.6e-08	0.90
8						
D.9009SA V1V2 Tags (LL)	0.52	100.0%	2.2	1.26 [1.05,2.30]	2.6e-08	1.05
A.9004SS V1V2 Tags (LL)	0.65	90.6%	1.7	0.79 [0.97,1.76]	2.6e-08	1.04
7						
E.A244 V1V2 Tags D7N (LL)	0.52	93.8%	2.2	0.71 [1.91,2.62]	2.6e-08	1.05
E.A244 V1V2 Tags (LL)	0.56	100.0%	1.3	0.48 [2.35,2.84]	2.6e-08	1.01
E.92TH023 V1V2 Tags (LL)	0.53	100.0%	1.8	0.95 [1.54,2.49]	2.6e-08	0.99
C.1086 V1V2 Tags (LL)	0.57	100.0%	2.3	0.60 [2.22,2.82]	2.6e-08	1.06
No Antigen						
No Antigen	0.34	6.2%	NA	0.02 [-0.01,0.01]	0.8	NA
6						
SIV mac239 J08 V1V2 (Nabel)	0.34	9.4%	5.8	0.15 [0.10,0.25]	1.8e-07	0.81
5						
C.J08 V1V2 (Nabel)	0.57	6.2%	3.2	0.14 [0.03,0.17]	4.9e-07	1.43
4						
B.J08 V1V2 (Nabel)	0.75	6.2%	2.6	0.08 [0.03,0.12]	3.1e-07	1.45
3						
A.gp70 A.Q23 V1V2 (LL)	0.75	12.5%	2.4	0.46 [0.25,0.71]	2.6e-08	0.99
B.gp70 CASE A2 V1V2 (LL)	0.89	62.5%	2.5	0.50 [0.23,0.72]	2.6e-08	0.15
C.gp70V1V2 (Nabel)	0.87	18.8%	3.7	0.44 [0.28,0.71]	2.6e-08	0.15
2						
B.gp70V1V2 (Nabel)	0.90	6.2%	3.0	0.45 [0.29,0.74]	5.2e-08	1.04
p565f pur. Gp70 V1V2 (Pinter)	1.00	6.2%	3.5	0.41 [0.29,0.70]	2.6e-08	0.00
1						
A.Q23 V1V2 Tags (LL)	0.79	15.6%	1.9	0.65 [0.34,0.99]	2.6e-08	1.02
B.63521 V1V2 Tags (LL)	0.80	9.4%	1.7	0.35 [0.24,0.59]	2.6e-08	0.94
B.CASE A2 V1V2 Tags (LL)	0.86	9.4%	2.1	0.40 [0.28,0.69]	1.7e-06	0.65

Table 3.1: Ten clusters based on baseline adjusted 1:100 dilution of the 20 V1V2-scaffolded proteins. Hierarchical clustering trees using complete linkage based on a dissimilarity measure defined as one minus the Spearman rank correlation coefficient between Week 26 readouts of vaccine recipients were used to cluster antigens into common antigenic groups. No antigen and MuLV are included in the table as clusters of size one but are not of primary interest. The columns shown in the table are: 1) Cluster, a numeric id for the cluster that separates each group of antigens; 2) Correlation with p565f, the Spearman correlation with the V1V2 antigen used in the RV144 correlates analysis; 3) Signal-to-noise ratio, the standard deviation of the subject specific readouts divided by the mean of the standard deviation of the within subject replicates; 4) Response Rate, based on a threshold defined as 3 standard deviations above the mean of the week 0 readouts; 5) IQR, difference between upper and lower quartiles followed by inter-quartile range in brackets; 6) P-value, P-value from a Wilcoxon rank sum test comparing readouts by treatment group; 7) Angle, angle between p565f and the specified antigen with vertex at the mediod of the plasma in an antigenic map of the V1V2 scaffold antigens and 32 vaccine plasma samples.

	Subtype	IC50 sCD4	GMT ID50	R-squared
246-F3 C10 2	AC (3.5%)	8.8 (50.0%)	51.3 (80.5%)	0.82
BJOX002000.03.2	CRF07 BC (5.5%)	1.1 (8.5%)	60.6 (87.5%)	
CE1176 A3	C (31.0%)	4.8 (31.0%)	51.5 (81.0%)	

Table 3.2: Three isolates selected using the lasso method described in Section 3.3.2. The columns show the subtype of each isolate and percentage of NSDP isolates of the same subtype; the IC50 of soluble CD4 (and percentile across all NSDP isolates); the geometric mean titer (GMT) of sera responses to the selected isolate (and percentile across all NSDP isolates); and the R-squared for the model using these three isolates.

	Subtype	IC50 sCD4	GMT ID50	R-squared
246-F3 C10 2	AC (3.5%)	8.8 (50.0%)	51.3 (80.5%)	0.95
249M B10	C (31.0%)	18.2 (78.5%)	29.9 (41.5%)	
7060101641A7(REV-)	C (31.0%)	0.8 (6.0%)	45.1 (68.5%)	
BJOX002000.03.2	CRF07 BC (5.5%)	1.1 (8.5%)	60.6 (87.5%)	
CE1176 A3	C (31.0%)	4.8 (31.0%)	51.5 (81.0%)	
CNE55	CRF01 AE (9.5%)	17.2 (76.5%)	25.4 (26.0%)	
HIV-25710-2.43	C (31.0%)	2.6 (17.0%)	110.1 (97.0%)	
TRO.11	B (25.5%)	11.5 (60.5%)	58.8 (87.0%)	
X1632 S2 B10	G (3.0%)	2.2 (15.5%)	47.4 (74.0%)	

Table 3.3: Nine isolates selected using the lasso method described in Section 3.3.2. The columns show the subtype of each isolate and percentage of NSDP isolates of the same subtype; the IC50 of soluble CD4 (and percentile across all NSDP isolates); the geometric mean titer (GMT) of sera responses to the selected isolate (and percentile across all NSDP isolates); and the R-squared for the model using these three isolates.

	Subtype	IC50 sCD4	GMT ID50	R-squared
234-F1-16-57	C (31.0%)	5.4 (35.5%)	56.4 (85.5%)	0.80
CNE8	CRF01 AE (9.5%)	50.0 (100.0%)	85.8 (95.0%)	
PRB931 06 TC3 4930	B (25.5%)	14.0 (68.0%)	35.2 (55.0%)	

Table 3.4: Three isolates selected using the tree distance clustering method described in Section 3.3.2. The columns show the subtype of each isolate and percentage of NSDP isolates of the same subtype; the IC50 of soluble CD4 (and percentile across all NSDP isolates); the geometric mean titer (GMT) of sera responses to the selected isolate (and percentile across all NSDP isolates); and the R-squared for the model using these three isolates.

	Subtype	IC50 sCD4	GMT ID50	R-squared
1105.P17.1	CRF02 AG (6.0%)	50.0 (100.0%)	36.1 (57.0%)	0.95
191821 E6 1	D (2.5%)	6.1 (38.5%)	29.7 (40.5%)	
234-F1-16-57	C (31.0%)	5.4 (35.5%)	56.4 (85.5%)	
CH119.10	CRF07 BC (5.5%)	11.3 (59.0%)	47.6 (74.5%)	
CNE8	CRF01 AE (9.5%)	50.0 (100.0%)	85.8 (95.0%)	
PRB931 06 TC3 4930	B (25.5%)	14.0 (68.0%)	35.2 (55.0%)	
Q842.D12	A (4.0%)	13.5 (66.5%)	35.6 (56.0%)	
SC422661.8	B (25.5%)	4.7 (30.5%)	30.9 (43.5%)	
X1632 S2 B10	G (3.0%)	2.2 (15.5%)	47.4 (74.0%)	

Table 3.5: Nine isolates selected using the tree distance clustering method described in Section 3.3.2. The columns show the subtype of each isolate and percentage of NSDP isolates of the same subtype; the IC50 of soluble CD4 (and percentile across all NSDP isolates); the geometric mean titer (GMT) of sera responses to the selected isolate (and percentile across all NSDP isolates); and the R-squared for the model using these three isolates.

Isolate	ALL	Q1	Q2	Q3	Subtype	Fiebig stage	EILU
246-F3 C10 2	1	1	1	1	AC	VI	I
249M B10	1	1	1	1	C	IV	E
CNE55	1	1	1	1	CRF01 AE	chronic	L
3728.V2.C6	1	1			C	III/IV	E
3103.V3.C10	1				ACD	V/VI	I
410-F2 1 30	1				C	VI	I
6101.10	1				B	V	I
CE703010217 B6	1				C	V/VI	I
HIV-16055-2.3	1				C	II	E
CE1176 A3		1	1	1	C	I/II	E
TRO.11		1	1	1	B	III	E
X1632 S2 B10		1	1	1	G	Chronic	L
C1080.C03		1			CRF01 AE	>VI	L
RPW-0510.2		1			B		U
7060101641A7(REV-)			1	1	C	I/II	E
BJOX002000.03.2			1	1	CRF07 BC	I/II	E
CH119.10			1		CRF07 BC	chronic	L
HIV-25710-2.43				1	C	V	I

Table 3.6: Nine isolate panels selected by the lasso/quartile method described in 3.3.6. A one in the All, Q1, Q2, or Q3 column indicates an isolate that was selected by the grouped lasso or quartile based regular lasso method respectively. The remaining columns classify the selected isolates by subtype, Fiebig stage and EILU (early, intermediate, late or unknown). The EILU categorization collapses across Fiebig stages to reduce the number of categories when identifying the time since infection.

Chapter 4

K-MER SCANNING

4.1 Introduction

Consider an HIV-1 vaccine efficacy trial. An important analysis compares HIV-1 sequences from breakthrough infections by treatment group. This type of analysis is known as a sieve analysis. From analyses of this type, we hope to better understand mechanisms of protection and/or post infection pressure that can be attributed to the vaccine. In this chapter we compare K-mer regions, small sets of amino acids, by treatment group. The typical K-mer is a short segment of contiguous amino acids, or peptide, of length 8 to 12. Peptides of this size are the most common length for a Cytotoxic T-Lymphocyte (CTL) epitope. It is well known that HIV-1 escapes CTL immune pressure by mutating in and around CTL epitopes and that these escape mutations are restricted by host HLA type. We focus on assessment of post infection pressure exerted via 9-mer CTL epitopes but note that the methods discussed in this chapter also apply to the setting of non-contiguous K-mers in the context pre-infection blocking effects due to antibody pressure.

The purpose of this chapter is to compare tests for assessing a treatment effect on HIV-1 sequences post-infection using marginal models for clustered data. We know HLA genotype is important in this setting but we defer treatment of it to the next chapter. In this chapter we introduce notational conventions in Section 4.2, provide a description of the statistical tests in Section 4.3, describe simulation methods in Section 4.4, show the results of simulations in Section 4.5, review bias correction methods in Section 4.6, explore multiplicity adjustment in Section 4.7, and finish with a discussion in Section 4.8. Following the discussion, a proof of the equivalence of two GEE bias adjustment procedures is provided in Section 4.9.

4.2 Notation

Consider n_p infections in the placebo arm and n_v infections in the treatment arm. Full genome HIV-1 sequences are collected from each infected subject at a time point that is typically between zero and six months post infection. The number of sequences collected for the i^{th} subject is a random variable M_i with observed value m_i. We can easily simulate missing data scenarios by allowing $m_i = 0$. For a fixed K-mer region, let $\Omega = \left\{\omega_1, \ldots, \omega_{|\Omega|}\right\}$

be the set of potential K-mers in this region. In theory this set has 21^K elements (since each of the K positions is one of 20 amino acids or the gap character) but in practice we consider the far smaller set of observed K-mers. For the i^{th} subject, given $M_i = m_i$, let $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{im_i})$ be a random vector of indices of the K-mers in Ω. The observed set of K-mers is specified by the vector of indices $\mathbf{y_i} = (y_{i1}, \ldots, y_{im_i})$. Although $\mathbf{Y_i}$ is a vector of integer valued variables that index the elements of Ω, we often think of it as a set of K-mers and will at times refer to it as the i^{th} subject's set of K-mers and similarly refer to $\mathbf{y_i}$ as the set of observed K-mers.

4.3 Statistical Methods: Hypothesis Testing and Estimation

Dating back to 1998 Gilbert and colleagues have been working on sieve analysis methods (e.g., [23, 20, 22, 24]). These methods either focus on a single sequence per subject or focus on scanning one site at a time. Applied sieve analyses by deCamp with Rolland et al. 2011, Gnanakaran et al. 2011, Janes et al. 2012, and Rolland et al. 2012 [71, 25, 36, 70] use novel methods for assessing multiple sequences per subject and collections of sites but the methodology has never been completely described and studied.

Here we extend the genome scanning approach of Gilbert, Wu, and Jobes (GWJ) [24] to handle both multiple sequences per subject and collections of sites. We pursue the K-mer scanning approach described in [71] with the goal of both refining the method and better characterizing its operating characteristics.

The K-mer scanning approach uses an HIV-1 PAM matrix [54] to generate a similarity score between each sequence and the vaccine insert strain. Specifically, for an alignment of the vaccine strain and a breakthrough strain, the K corresponding entries in the PAM matrix are summed to generate a similarity score; to account for gaps we add a row and a column to the PAM matrix with entries equal to the smallest value in the matrix. We are interested in inferences on the treatment effect defined as the difference in marginal mean similarity scores between treatment groups. Generalized estimating equations (GEEs) [45, 87] were developed to fit marginal models to data of this type where within a cluster (i.e. subject), the outcomes (i.e. similarity scores) are correlated. In our use of GEE, each cluster is defined by an infected subject from an HIV vaccine trial with one similarity score

to the vaccine insert reference strain for each breakthrough sequence. In our analysis we are interested in the issue of informative cluster size (i.e., when the distribution of the outcome within a cluster depends on the cluster size) since when cluster size is informative GEE can give biased results.

We are particularly interested in marginal methods that remain valid when cluster size is informative since this may be the case in an analysis of post-infection vaccine effects. In 2001, Hoffman et al. [31] introduced a Monte Carlo method for analyzing clustered data with informative cluster size called within-cluster resampling (WCR). Follmann et al. [17] further develop the WCR method under the moniker multiple outputation (MO); subsequently we refer to within-cluster resampling using the Follman terminology MO. At roughly the same time Williamson et al. [84] developed cluster-weighted generalized estimating equations (CWGEE) to handle data with informative cluster size.

In the following treatment we fit models using all three methods; GEE, MO, and CWGEE. All three methods require a marginal model and the GEE methods require a model for within cluster correlation, a working correlation, while MO implicitly handles within cluster correlations. Although our data are inherently discrete, with typically between 6 and 12 distinct K-mers at a given position, we use a marginal Gaussian model. Even for discrete data, marginal Gaussian models provide consistent estimators; and, by using the PAM similarity matrix to generate scores for each K-mer, we put the outcome on a scale that has a continuous interpretation. We could also consider marginal models for discrete data but do not pursue this approach further. The main reason is that when scanning a protein we may have K-mer regions that are particularly amenable to a discrete method but we believe that fit issues are likely to arise when trying to apply discrete methods to regions that have rare K-mer variants. For within cluster correlations in the GEE and CWGEE models, we use an exchangeable working correlation.

Since these methods are used to scan a protein for regions with a treatment effect, they are applied to multiple overlapping K-mers. Therefore, multiplicity adjustment procedures are required to interpret the results and we are particularly interested in multiplicity adjustment procedures that can handle correlated outcomes. We are interested in both family wise error rate (FWER) and false discovery rate (FDR) procedures. Several groups have

considered multiplicity adjustment for correlated outcomes; for example, van der Laan, Dudoit, and Pollard 2004 [79], Conneely and Boehnke 2007 [10], and Efron 2007 [13]. In this chapter, we extend P_{ACT} FWER adjustment procedure of [10] to this application. We also explore extending the P_{ACT} to the FDR framework. Furthermore, we use adjustment procedures that do not account for between test correlations; specifically, Holm's adjustment [32] for FWER and the Benjamini and Hochberg adjustment [3] for FDR.

4.4 Simulation Methods

We wish to simulate data that is comparable to the Step vaccine trial. Initially, we consider a fixed number of infections in an HIV-1 efficacy trial with 26 infections in the placebo group and 39 infections in the vaccine group; the same number of subjects available for study of the Gag protein in the Step trial sieve analysis [71]. Each infected subject had between 1 and 12 sequences available for analysis with a mean of approximately 6.5 sequences. We consider two methods for data generation. The first method simulates multiple sequences per subject for a single 9-mer location using a single parameter ρ to parameterize the amount of within subject correlation. The second method simulates entire sequences using two parameters, ρ_1 and ρ_2 to parameterize the within subject correlations both within and between 9-mer positions.

4.4.1 Simulation Method 1

For this method we assume a parametric form of the joint distribution $(M_i, \mathbf{Y_i})$. We assume that the number of sequences M_i ranges from 1 to 12 per subject following a multinomial distribution with probabilities $\mathbf{p} = (p_1, ..., p_{12})$ where the probabilities are derived from a mean λ Poisson distribution, P_λ. We define

$$\mathbf{p} = (\Pr(P_\lambda \leq 1), \Pr(P_\lambda = 2), \ldots, \Pr(P_\lambda = 11), \Pr(P_\lambda \geq 12)). \tag{4.1}$$

Typically investigators will try to get about 10 sequences per subject and will go to great lengths to sequence at least one virus from every infected individual. We simulate the

number of sequences per subject seen in Step[1] using $\lambda = 6.5$.

Next, we simulate a set of K-mer peptides for each subject, $\mathbf{Y_i}$, by selecting indices from an observed set of unique K-mers, Ω, with observed frequencies $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{|\Omega|})$. We set the amount of within subject correlation between K-mers using a correlation parameter ρ. We want the members of Ω sensibly ordered (e.g. by similarity to reference strain under some metric) since ρ will imply that, within a subject, observed K-mers will have similar ranks under the ordering metric. This is a reasonable assumption when working with subjects that were infected with a single founder sequence[2] in which case we expect K-mers found within a subject to be related.

Given $M_i = m_i$, we generate $\mathbf{Y_i}$ by transforming an m_i-variate random normal variable,

$$\mathbf{Z_i} \sim \mathcal{N}_{m_i}(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is an exchangeable correlation matrix with correlation ρ,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}_{m_i \times m_i}.$$

The vector $\mathbf{Z_i}$ is transformed into a m_i-variate uniform sample, $\mathbf{U_i}$, using the inverse standard normal CDF, ϕ^{-1},

$$\mathbf{U_i} = (\phi^{-1}(Z_{i1}), \ldots, \phi^{-1}(Z_{im_i})).$$

Next, we partition the unit interval into $|\Omega|$ sections using cut points defined by $\boldsymbol{\pi}$,

$$\boldsymbol{\xi} = (\pi_1, \pi_1 + \pi_2, \ldots, \sum_{j=1}^{|\Omega|-1} \pi_j, 1). \tag{4.2}$$

Finally, we define the j^{th} element of $\mathbf{Y_i}$, Y_{ij}, by

$$Y_{ij} = \underset{1 \le l \le |\Omega|}{\operatorname{argmin}} \{U_{ij} \le \xi_l\}.$$

---

[1] In the Step trial, counts of 5 and 10 sequences are very common so this form of data generation is not particularly true to the observed sequence counts in Step.

[2] To simulate data for a subject with multiple founders we can simulate the sequences as though they were obtained from independent observations.

Notice that this definition makes Y_{ij} an integer valued random variable with support equal to the integers from 1 to $|\Omega|$. As previously mentioned, we can also think of Y_{ij} as a K-mer valued categorical random variable since it indexes the set Ω.

Under this simulation setup, the marginal probability of a given K-mer is maintained in that $\Pr(Y_{ij} = k) = \pi_k$, the observed frequency of the k^{th} K-mer in Ω. Since we used an exchangeable correlation matrix, the correlation between K-mers within a subject $\rho_{\text{within}} = \text{Cor}(Y_{ik}, Y_{il})$ is constant over all $k \neq l$. Simulations show that ρ_{within} is attenuated compared to the data generating correlation ρ due to the transformation to a uniform random variable as well as the discretization process.

Specifying Alternatives with Non-informative Cluster Size under Simulation Method 1

For a fixed K-mer region, assume that the K-mers have been ordered by some metric when compared to the reference vaccine insert sequence so that ω_1 is closest to the reference and $\omega_{|\Omega|}$ is furthest from the reference. To produce a difference between treatment groups we can simulate a pattern of escape away from reference strain by either 1) using different Ωs and associated $\boldsymbol{\pi}$s for each treatment group or, 2) skewing the set of $\boldsymbol{\pi}$s towards ωs that are farther from the reference. We choose the second approach since we can parametrize the skew if we exponentiate the cut points $\boldsymbol{\xi}$.

For the placebo group, generate Y_i using $\boldsymbol{\xi}$ from equation 4.2. For the vaccine group use,

$$\boldsymbol{\xi}_v = (\xi_1^\epsilon, \ldots, \xi_{|\Omega|}^\epsilon) \tag{4.3}$$

where $\epsilon > 1$. If we define $\xi_0 = 0$, then, given the vector of cut points in equation 4.2, the probabilities are given by

$$\boldsymbol{\pi} = \left(\xi_1 - \xi_0, \xi_2 - \xi_1, \ldots, 1 - \xi_{|\Omega|-1}\right). \tag{4.4}$$

Similarly, for the cut points in equation 4.3, the probabilities are given by

$$\boldsymbol{\pi}_v = \left(\xi_1^\epsilon - \xi_0^\epsilon, \xi_2^\epsilon - \xi_1^\epsilon, \ldots, 1 - \xi_{|\Omega|-1}^\epsilon\right). \tag{4.5}$$

The transformation of the cut points given in equation 4.2 to those given in equation 4.3

induces a transformation of the probabilities $\boldsymbol{\pi}$ in equation 4.4 to the probabilities $\boldsymbol{\pi}_v$ in equation 4.5. The transformation has the geometric interpretation described below.

Given the polynomial $f(x) = x^{\epsilon}$, $\pi_{v,j}$ is π_j multiplied by the slope of the secant line through the points $(\xi_{j-1}, f(\xi_{j-1}))$ and $(\xi_j, f(\xi_j))$. The derivative of f is 1 at $\xi^* = \epsilon^{\frac{1}{1-\epsilon}}$, thus when $\xi_j \leq \xi^*$, $\pi_{v,j} < \pi_j$ and when $\xi_{j-1} \geq \xi^*$, $\pi_{v,j} > \pi_j$. For each successive probability $\pi_{v,j}$, $j = 1, \ldots, |\Omega|$, the associated secant line becomes steeper. If the original probabilities, $\boldsymbol{\pi}$, are all greater than zero and $|\Omega| \geq 2$, then at least one slope will be less than one and one slope will be greater than one, so the probabilities are scaled smaller at first then bigger once $\xi_{j-1} \leq \xi^* < \xi_j$. The larger the value of ϵ, the more skewed the distribution of Y_{ij} towards K-mers that are less like the reference.

As an example (Figure 4.1) consider a hypothetical K-mer region. Suppose that in this region $\Omega = (\omega_1, \omega_2, \omega_3)$ where ω_1, ω_2, and ω_3 are equally likely K-mers with 0, 1, and 2 amino acid mismatches, respectively, when compared to the reference K-mer. Thus the probabilities of observing the elements of Ω in the placebo group are given by $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$ with the cut points defined by $\boldsymbol{\xi} = (1/3, 2/3, 1)$. For $\epsilon = 2$ we get $\boldsymbol{\xi}_v = (1/9, 4/9, 1)$ which can be converted to probabilities on Ω for the vaccine group of $\boldsymbol{\pi}_v = (1/9, 1/3, 5/9)$ demonstrating the skew towards K-mers that have a greater dissimilarity with the reference.

Specifying Alternatives with Informative Cluster Size under Simulation Method 1

An important alternative for assessing K-mer scanning methods will be to consider informative cluster size. Here we consider the scenario where immune pressure is a function of viral load. We assume that a lower viral load is associated with a smaller number of viruses sequenced and that in the vaccine group a lower viral load is associated with more immune pressure in a direction away from the reference. To simulate data with informative cluster size we let $\boldsymbol{\xi}_v$ vary with m_i using the following formula,

$$\boldsymbol{\xi}_v = \begin{cases} (\xi_1^{\epsilon}, \ldots, \xi_{|\Omega|}^{\epsilon}), & \text{if } m_i > 3 \\ (\xi_1^{\eta \cdot \epsilon}, \ldots, \xi_{|\Omega|}^{\eta \cdot \epsilon}), & \text{if } m_i \leq 3 \end{cases}$$

where $\epsilon \geq 1$ and $\eta > 1$ is a multiplicative factor that simulates extra immune pressure in small clusters.

Let T be an indicator of treatment status, with 1 indicating active treatment and 0 indicating placebo. We could simulate data under which $M_i|T = 1$ and $M_i|T = 0$ have different distributions but we consider the simpler case described above where the distribution of M_i does not depend on treatment status.

4.4.2 Simulation Method 2

The second method of data simulation simulates complete Gag sequences based on sampling Gag sequences from the infected placebo group. To motivate the simulation plan for full length sequences we first consider the data available. The purest simulation does not mix sequences from the vaccine and placebo groups thus we are limited to sequences from the 26 infected placebo recipients with Gag sequence data. In order to replicate independent data for a trial with 26 placebo recipients and 39 vaccine recipients we first consider the diversity seen in the Gag sequences from the placebo recipients. The mean and the standard deviation of the between subject hamming distances are roughly 50 and 10 amino acid differences respectively. For within subject diversity the mean hamming distance between sequences is only about 1.5. Our goal is to mimic a set of sequences with similar amounts of within and between subject diversity.

We first sample 65 sequences with replacement from the set of 153 available Gag sequences from the infected placebo group. These 65 sequences are intended to represent independent sequences but we will necessarily sample multiple sequences from the same placebo subject. In order to simulate independent sequences, we mutate each sequence such that the mean pairwise hamming distance between sequences is approximately 50 amino acid differences with a standard deviation of about 10. To achieve this, we randomly select 50 alignment positions, separately for each selected sequence, and "mutate" the corresponding amino acid at each position according to the observed frequency of amino acids at that position. Use Θ to denote this set of sequences.

Once we have simulated a random sample of one sequence per subject in this way, we simulate multiple sequences per subject in a manner similar to that used to generate correlated K-mers in Method 1. The added complication is that not only do we want

the amino acid distribution at each site to be correlated, we also want the amino acid distributions between sites to be correlated.

To accomplish this consider the joint distribution $(M_i, \mathbf{Y_i})$ where, given $M_i = m_i$, $\mathbf{Y_i}$ is a vector of m_i multiply aligned sequences. We will not fully describe the distribution of $\mathbf{Y_i}$, although one could. Rather, we describe the data generation method which "mutates" each sequence in the of 65 independent, multiply aligned Gag sequences denoted by Θ which were derived by the process described above. We first choose a vector of positions from those alignment positions where the consensus amino acid represents less than 90% of the sequences. For these data, this rule selects 100 sites. For each subject i, we sample E_i of these sites where $E_i \in \{1, 2, \ldots, 18\}$. Eighteen is slightly larger than the upper limit of the number of variable sites seen in the breakthrough Gag sequences in the placebo group. Our approach will not guarantee that a selected mutation site has variation, so we use an upper limit of 18 which results in roughly the same number of variable sites in the observed sequences. We generate observed values e_i using a Poisson distribution P_λ, with mean $\lambda = 11$, using a variation of equation 4.1,

$$\mathbf{p} = (\Pr(P_\lambda \leq 1), \Pr(P_\lambda = 2), \ldots, \Pr(P_\lambda = 17), \Pr(P_\lambda \geq 18)).$$

Given this vector of positions, $\boldsymbol{\gamma}$, where $|\boldsymbol{\gamma}| = e_i$, we mutate the $m_i \cdot e_i$ amino acids for a given subject using a transformation similar to the one used in Method 1.

We first generate an $m_i \cdot e_i$-variate random normal distribution

$$\mathbf{Z_i} \sim \mathcal{N}_{m_i \cdot e_i}(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is an $e_i \times e_i$ block form matrix with blocks of $m_i \times m_i$ exchangeable correlation matrices on the diagonal and $m_i \times m_i$ block constant matrices elsewhere.

$$\boldsymbol{\Sigma} = \begin{bmatrix} A & B & B & \cdots & B \\ B & A & B & \cdots & B \\ B & B & A & \cdots & B \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B & B & B & \cdots & A \end{bmatrix}_{m_i \cdot e_i \times m_i \cdot e_i} \quad \text{with}$$

$$A = \begin{bmatrix} 1 & \rho_1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & \rho_1 & 1 & \cdots & \rho_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \rho_1 & \cdots & 1 \end{bmatrix}_{m_i \times m_i} \quad \text{and}$$

$$B = \begin{bmatrix} \rho_2 & \rho_2 & \rho_2 & \cdots & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & \cdots & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & \cdots & \rho_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_2 & \rho_2 & \rho_2 & \cdots & \rho_2 \end{bmatrix}_{m_i \times m_i} .$$

As mentioned above, we will use $\rho_1 > \rho_2$.

Again, we proceed in a manner similar to the approach in Method 1. The only difference is that we need to use different cut points for the different alignment positions. These cut points, given by $\boldsymbol{\xi}_k$, are determined by the observed amino acid distribution at alignment position γ_k in the set of Gag sequences from all placebo group breakthrough infections.

The vector $\mathbf{Z_i}$ is transformed into an $m_i \cdot e_i$-variate uniform sample, $\mathbf{U_i}$, using the inverse standard normal CDF, ϕ^{-1},

$$\mathbf{U_i} = (\phi^{-1}(Z_{i1}), \ldots, \phi^{-1}(Z_{im_i \cdot e_i})).$$

For alignment position k, we index the observed amino acids in the set of Gag sequences from all placebo group breakthrough infections at that position and, given subject i and sequence j, we define $Y_{ij}(k)$ by

$$Y_{ij}(k) = \underset{l}{\text{argmin}}\{U_{ij} \leq \xi_{kl}\}.$$

where l ranges over the integers from 1 to the number of unique amino acids at the k^{th} alignment position, γ_k.

4.5 Simulation Results

4.5.1 Simulation Method 1 Results

We simulated a null distribution at SLYNTVATL (HXB2 positions 77 to 85) and KNCRAPRKK (HXB2 positions 403 to 411) in the Gag protein under simulation Method 1. SLYNTVATL was chosen as a well known HLA-A*02 restricted epitope that is encoded by the vaccine used in the Step trial. KNCRAPRKK was chosen to represent a relatively low diversity K-mers region. We simulated 10,000 data sets of 26 placebo and 39 vaccine recipients under the null hypothesis of no treatment effect using a correlation of $\rho = 0.9$. The observed Type I error rate for $\alpha = 0.05$ for each test at each 9-mer location are shown in Table 4.1. MO controls the size of the test at the nominal rate while GEE is slightly over size.

Next, we simulate a non-informative cluster size alternative using Method 1 with $\epsilon = 2.5$ and $\rho = 0.9$. The power of both tests to detect a treatment effect are nearly identical for both tests across both simulated 9-mer regions as shown in Table 4.2. Both tests appear to provide unbiased estimates of the true treatment effect; however, coverage of the true parameter is slightly below the nominal level for a 95% confidence interval as shown in Table 4.3.

Finally, we simulate an informative cluster size alternative using Method 1 with $\epsilon = 2$, $\eta = 5$, and $\rho = 0.1$. It was necessary to use a correlation parameter ρ close to zero in order to see a biased result in the GEE model. This is likely because high values for ρ greatly diminish the information carried in the cluster size. As expected, the MO estimates appear to be unbiased (see Table 4.4). Coverage of the true parameter is slightly below the nominal rate again for the MO test under this alternative while the coverage for GEE is very low as was expected since we have an informative cluster size alternative (see Table 4.5).

4.5.2 Simulation Method 2 Results

For simulation Method 2 we consider only the GEE test under the complete null hypothesis of no treatment effect at any K-mer region. To test the complete null hypothesis for GEE we simulate 1,000 trials of 26 placebo and 39 vaccine breakthrough infections as described

in Section 4.4.2. We then scan the 534 9-mer regions for a treatment effect using our GEE based test. We compute the minimum adjusted P-value for each simulation using the Holm FWER multiplicity adjustment procedure (Figure 4.2). The Holm adjustment is known to be conservative for correlated tests. Interestingly, we see that the FWER is controlled almost exactly at the nominal rate when $\alpha = 0.05$. At nominal rates greater than $\alpha = 0.05$ the Holm method is clearly conservative.

4.6 Bias corrections to the sandwich variance estimator for small samples

The sandwich estimator for the variance of the regression parameters often used in GEE is consistent even when the correlation model is mis-specified but the estimator is biased (see Theorem 5.4 of [88]). This bias tends to underestimate the variance which leads to under coverage of confidence intervals and inflated type I errors particularly in small sample settings. As seen in earlier simulations of the Step trial setting, bias in the sandwich variance estimate is large enough to raise concerns regarding coverage and type I error rates and these problems will only be accentuated when doing subgroup analyses.

Numerous studies have been conducted and several solutions have been proposed to this issue of bias (e.g. see [15, 48, 58, 39, 59, 52, 47, 82] and reviews [88, 11]). The proposed solutions generally fall within three categories. The first solution is to abandon the use of the sandwich estimator and resort to either a Jackknife or a bootstrap estimator of the variance. This option will not be considered further. A second solution is to use some form of bias correction with suggestions given in [15, 48, 39, 52]. A third solution is to take into account the variability of the sandwich estimator and base inferences on a t- or F-test; an idea which parallels studentizing the sample mean [59, 39]. Combinations of the second and third solutions have been proposed but in this exploration of small sample corrections we consider them separately.

4.6.1 Generalized estimating equations and the sandwich estimator of variance

Assume we have a vector of responses $Y_i = (Y_{i1}, \ldots, Y_{iT_i})^T$ from n subjects with T_i outcomes for the i^{th} subject. For each response Y_{it} we have a p-dimensional covariate vector X_{it} therefore for each subject we have the $T_i \times p$ covariate matrix $X_i = (X_{i1}, \ldots, X_{iT_i})^T$.

The marginal regression model relates the mean $E(Y_{it}|X_{it}) = \mu_{it}$ to the covariate vector X_{it} via the generalized linear model $g(\mu_{it}) = X_{it}\beta$ where g is a known link function and β is a p-dimensional vector of regression coefficients. As with generalized linear models we write the variance as $\nu_{it} = V(Y_{it}|X_{it}) = \phi \cdot h(\mu_{it})$, where ϕ is the dispersion parameter and h is the variance function. Let $\Omega_i = V(Y_i|X_i)$ denote the conditional $T_i \times T_i$ covariance matrix. The GEE framework allows us to specify a working covariance matrix Σ_i in place of the true covariance matrix Ω_i. Given an estimate $\hat{\Sigma}_i$ of the working correlation matrix Σ_i, we estimate β by solving the estimating equations

$$\sum_{i=1}^{n} D_i^T \hat{\Sigma}_i^{-1}(Y_i - \mu_i) = 0$$

where $\mu_i = (\mu_{i1}, \ldots, \mu_{iT_i})^T$ and $D_i = \partial\mu_i/\partial\beta^T$. The covariance matrix $V(\hat{\beta})$ of $\hat{\beta}$ is consistently estimated by the sandwich estimator of variance. The sandwich variance matrix is given by

$$V(\hat{\beta}) = \frac{1}{n} A^{-1} B A^{-1} \tag{4.6}$$

with

$$A = \textstyle\sum_{i=1}^{n} D_i^T \Sigma_i^{-1} D_i \quad \text{and} \quad B = \sum_{i=1}^{n} D_i^T \Sigma_i^{-1} \Omega_i \Sigma_i^{-1} D_i \ .$$

When the working covariance is correctly specified, the sandwich variance matrix of equation 4.6 simplifies to the Fisher information matrix A^{-1}. Whereas the Fisher information matrix is the bread of equation 4.6 the cheese portion, B, is called the outer product gradient (OPG). For estimation, residuals $Y_i - \mu_i$ are used to estimate Ω_i which is replaced in equation 4.6 by $\hat{\Omega}_i = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T$.

As shown in Mancl and DeRouen [48],

$$E[\hat{\Omega}] \approx (I_i - H_{ii})\Omega_i(I_i - H_{ii})^T + \sum_{i \neq j} H_{ij}\Omega_j H_{ij}^T, \tag{4.7}$$

where $H_{ij} = D_i A^{-1} D_j^T \Sigma_j^{-1}$ and I_i is a $T_i \times T_i$ identity matrix.

Bias correction methods

We consider four different bias correction methods.

1. Based on the expectation shown in equation 4.7 and the assumption that the sum in that equation is negligible, Mancl and DeRouen [48] propose the following adjustment to the usual OPG estimator

$$\hat{B}_{MD} = \sum_{i=1}^{n} \hat{D}_i^T \hat{\Sigma}_i^{-1} (I_i - \hat{H}_{ii})^{-1} \hat{\Omega}_i (I_i - \hat{H}_{ii})^{-1} \hat{\Sigma}_i^{-1} \hat{D}_i$$

2. Kauermann and Carroll [39], propose adjusting the OPG estimator using

$$\hat{B}_{KC} = \sum_{i=1}^{n} \hat{D}_i^T \hat{\Sigma}_i^{-1} (I_i - \hat{H}_{ii})^{-1/2} \hat{\Omega}_i (I_i - \hat{H}_{ii})^{-1/2} \hat{\Sigma}_i^{-1} \hat{D}_i.$$

   Kauermann and Carrol assume that Σ_i is correctly specified while Mancl and DeRouen do not, which accounts for the different derivations. As noted by Lu et al. [47], the elements of H_{ii} are between 0 and 1, so $\hat{B}_{KC}$ are expected to give larger standard errors than the usual OPG estimator, $\hat{B}_{OPG}$, and $\hat{B}_{MD}$ larger than $\hat{B}_{KC}$.

3. We consider the modified Fay and Graubard bias correction of Ziegler [88] which is given by

$$\hat{B}_{mFG} = \sum_{i=1}^{n} \tilde{H}_i \hat{D}_i^T \hat{\Sigma}_i^{-1} \hat{\Omega}_i \hat{\Sigma}_i^{-1} \hat{D}_i \tilde{H}_i^T,$$

   where $\tilde{H}_i = (I - \hat{A}_i \hat{A}^{-1})^{-1/2}$, I is the $p \times p$ identity matrix and $\hat{A}_i = \hat{D}_i^T \hat{\Sigma}_i^{-1} \hat{D}_i$.

4. The proposed correction of Morel, Bokossa, and Neerchal is fundamentally different than the three earlier proposed corrections in that they use an additive correction term. Let $\hat{V}(\hat{\beta})$ be the usual plug in sandwich estimator using $\hat{B}_{OPG}$. Then

$$\hat{V}_{MBN}(\hat{\beta}) = \hat{V}(\hat{\beta}) + \hat{\delta}_n \cdot \hat{\zeta} \cdot \hat{A}^{-1},$$

   where $\hat{\delta}_n = \min\left(\frac{1}{2}, \frac{p}{n-p}\right)$ and $\hat{\zeta} = \max(1, \text{trace}(\hat{A}^{-1}\hat{B}_{OPG})/p)$.

As shown in Section 4.9 the KC and mFG are equivalent adjustment procedures. The square root of the inverse term in KC is $T_i \times T_i$ and for mFG it is $p \times p$. Thus the choice of implementation might be driven by the size of the typical cluster versus the number of covariates in the model; a smaller matrix would likely result in more stability in the numeric computation of the adjustment factor.

Accounting for variability of the sandwich estimator

As an alternative to the bias correction solutions described in the previous section, Pan and Wall (PW) [59] propose using an approximate t-test that takes into account the variability of the OPG. Pan and Wall first define the operator vec$(\cdot)$ that operates on matrices. Given a matrix U, vec(U) is a vector formed by stacking the columns of U below one another. They derive an estimator for the covariance of vec(B) where B is the OPG of equation 4.6. The estimator is

$$\widehat{\text{cov}}(\text{vec}(B)) = n^2(A \otimes A)\hat{V}(A \otimes A), \tag{4.8}$$

where $\hat{V} = \frac{1}{n(n-1)} \sum_{i=1}^{n}(b_i - \bar{b})(b - \bar{b})^T$, $b_i = \text{vec}(D_i^T \Sigma_i^{-1} \hat{\Omega}_i \Sigma_i^{-1} D_i)$, and $\bar{b} = \sum_{i=1}^{n} b_i/n$. From equation 4.8, we have the estimated variance of B_{kk}, the k-th diagonal element of B. Using a scaled chi-squared distribution $c\chi_d^2$, PW approximate the distribution of B_{kk}. Matching the first to moments of B_{kk} and $c\chi_d^2$, PW obtain

$$c = {}^{\tau_k}/_{2\sigma_k} \quad \text{and} \quad d = {}^{2\sigma_k^2}/_{\tau_k}$$

,

where τ_k and σ_k are the estimated mean and variance of B_{kk}, respectively. Since under $\text{H}_0 : \beta_k = 0$, ${}^{\hat{\beta}_k}/_{\sqrt{\sigma_k}}$ is approximately $N(0,1)$, and B_{kk}/c is approximately χ_d^2, then

$$t = \frac{{}^{\hat{\beta}_k}/_{\sqrt{\sigma_k}}}{\sqrt{B_{kk}}/cd} = \frac{\hat{\beta}_k}{\sqrt{B_{kk}}}$$

can be approximated by a t-distribution with $d = {}^{2\sigma_k^2}/_{\tau_k} \approx {}^{2B_{kk}^2}/_{\tau_k}$ degrees of freedom, where an estimate of τ_k is given by the k-th diagonal element of equation 4.8.

Comparison of bias correction methods

Theorem 5.21 of Ziegler [88] shows that the Mancel and DeRouen (MD) and modified Fay and Graubard (mFG) estimators are less biased than the usual sandwich variance estimator. This proof extends to the Kauermann and Carroll (KC) estimator since mFG and KC are equivalent. We see empirically that the variance estimate of MD is always greater than that of mFG and as discussed above we expect analytically that the variance estimate of MD is greater than that of KC which is confirmed in simulations. The adjustment of Morel,

Bokossa, and Neerchal (MBN) is additive and always positive so the Type I error is always smaller than that of the unadjusted sandwich estimator. In simulations, inference based on MBN tends to be less conservative than inference based on MD but more conservative than inferences based on mFG and KC. The Pan and Wall (PW) estimator is guaranteed to have a reduced Type I error rate compared to the usual sandwich estimate [59]. Empirically we see that inferences based on PW estimator can be either more conservative or more liberal than MD, mFG, KC, and MBN depending on the setting.

To summarize, we know that all five bias correction methods discussed here are guaranteed to have better Type I error control than the standard sandwich estimator. Analytical and empirical evidence suggest that listed from most conservative to least conservative we have MD, MBN, then mFG and KC. For PW, Type I error control relative to the other methods depends on the setting.

4.6.2 Simulation study of sandwich estimator bias corrections.

Simulated null data sets using simulation method 1 (see Section 4.4.1) were generated for K-mer regions SLYNTVATL and KTLRAEQAS using $\rho = 0.9$ under three different sample size scenarios. The SLYNTVATL region represents a diverse K-mer location, based on breakthrough sequences from the Step sieve study, with approximately 85% of K-mer regions in the Gag protein having lower entropy. The KTLRAEQAS region was chosen to represent low diversity with approximately 33% of K-mer regions in Gag having lower entropy. Point estimates and 95% confidence intervals of the level 0.05 Type I error rate are shown for the MD, mFG, KC, MBN, PW and standard sandwich estimators based on 10,000 data simulations for each sample size and K-mer region. The sample sizes are intended to represent 1) a subgroup analysis, 2) a trial setting with a moderate number of infections and 3) a trial setting with a large number of infections. The first sample size is based on the number of A02 positive participants from the Step sieve cohort with 11 placebo and 17 vaccine recipients. The second sample size is the entire Step sieve cohort with 26 placebo and 39 vaccine recipients. The third sample size is based on the Vax004 sieve cohort with 119 placebo and 217 vaccine recipients.

Figure 4.3 shows the observed Type I error rate and 95% confidence intervals for the six methods described above based on a level $\alpha = 0.05$. The unadjusted sandwich estimator is biased in the smaller sample sizes but attains nominal Type I error rate at the largest simulated sample size. This bias appears to be worst in the smallest sample size as well as under the lower diversity simulation. All methods do reasonably well at attaining nominal Type I error rate given the largest simulated sample size so we focus comparisons of the remaining five methods on the two smaller simulated sample sizes. As noted above, there is a consistent relationship between the MD, MBN, mFG, and KC estimators across sample sizes regardless of the diversity. The relationship is that the Type I error rate is lowest for MD and highest for mFG and KC with identical results and MBN lies between the other three. Notice that the PW estimator is conservative in the low diversity setting and anti-conservative for the high diversity setting. Across all simulations the MD methods appear to best approximate the nominal Type I error.

4.7 Multiplicity Adjustment

As noted earlier, multiplicity adjustment for correlated tests is important for interpreting the results of the K-mer scanning procedure. Here we extend the work of Conneely and Boehnke 2007 [10] to provide a FWER adjustment procedure in the context of K-mer scanning. In their work, the proportion of tests (L) to independent observations (N) is quite low and they recommend 0.1 as the upper limit for this ratio (i.e., $L \leq N/10$). They also generally have a large sample size with $N > 1,000$. In our application, we are likely to encounter $L >> N$ and $N \leq 100$. The Conneely and Boehnke 2007 FWER adjustment method P_{ACT} requires estimation of the covariance structure of the test statistics. Their method uses the sample covariance which, under scenarios studied in their work, has up to approximately 500 parameters and can be effectively estimated since the sample size is large. In our scenario the sample size is quite small compared to the number of parameters in the covariance matrix which can approach a half a million for a gene with 1,000 amino acids.

It is well known that for a small sample size a high-dimensional sample covariance matrix is ill-conditioned, that is it has a large ratio of largest to smallest eigen values, and

is often singular. Many alternatives to the sample covariance matrix have been proposed (see citations in Won and Kim [85]). Here we consider the method of Ledoit and Wolf [43] which shrinks the sample covariance S towards a matrix F with fewer parameters using a scale factor $\hat{\delta}$ (i.e., $\hat{\Sigma} = \hat{\delta}F + (1 - \hat{\delta})S$). Efron 2007 [13] uses a similar approach in his work on FDR adjustment for correlated tests by considering a single measure of correlation for the order statistics of the tests.

The method of Conneely and Boehnke 2007 [10] can be applied in our scenario of $L << N$ but the method is anti conservative. In a simulation study of quantitative traits with a correlation $\rho = 0.7$, $N =$ 2,000 and 600, 750, or 1,000 tests they note slightly inflated FWER. In our own two-sample simulation test of normally distributed outcomes, using $\rho = 0.7$, $N = 30$ and 10, 30, or 100 tests we also see an inflated FWER under the complete null hypothesis when the number of tests is large compared to the sample size. In our simulations the observed FWER for $\alpha = 0.05$ was 0.0483, 0.0562, and 0.0695 respectively for 10, 30, and 100 tests.

Within the framework of P_{ACT} we replace the sample correlation matrix with a regularized correlation matrix using the method of Ledoit and Wolf [43] and refer to this modified version of P_{ACT} as P_{rACT}. We use the fixed matrix F which consists of the sample variances along the diagonal and the average between test covariance in the off-diagonal positions. In a simulation study, we consider 30 subjects with 100 correlated outcomes. Here, rather than using a single between outcome correlation, we consider a more complicated correlation structure. Briefly, a 100 by 100 correlation matrix was generated with an average between outcome correlation of approximately 0.13 and a standard deviation of the between outcome correlations of approximately 0.17. Using this correlation matrix we simulated data under the complete null hypothesis. The FWER for $\alpha = 0.05$ was 0.0576 and 0.0481 for P_{ACT} and P_{rACT} respectively. The 95% confidence intervals for these error estimates exclude α for the P_{ACT} method while it includes α when for P_{rACT}.

The examples above are based on a single outcome per subject but we would like to apply the P_{rACT} method when we have multiple readouts per subject. For the MO method, Follmann, Proschan, and Leifer 2003 [17] provide the following estimator for the covariance matrix of the vector of MO estimates, $\hat{\theta}$, with an estimate of the associated covariance

matrix $\hat{\Sigma}$

$$\widehat{\text{var}}_{MO}(\bar{\hat{\theta}}) = \sum_{k=1}^{M} \frac{\hat{\Sigma}_k}{M} - S^2_{\hat{\theta}|\underline{\mathbf{X}}}, \tag{4.9}$$

where $S^2_{\hat{\theta}|\underline{\mathbf{X}}}$ is the sample covariance of the $\hat{\theta}$'s and M is the number of outputations. The simulation results described above suggest that when applying P_{ACT} multiplicity adjustment we should regularize the covariance matrix of equation 4.9. We ran two simulation scenarios, the first with 10 tests and the second with 25 tests. In both cases we used $N = 50$ subjects, 5 replicates per subject, a between test correlation of 0.2 and a within subject correlation of 0.8. We used $M = 2{,}000$ outputations to generate the covariance matrix of equation 4.9. One problem that arises, and seems likely to depend on both the dimension as well as the number of outputations used to generate the matrix defined by equation 4.9, is that the matrix is not always positive definite. For the simulation with 10 tests, 16% of our simulations did not result in a positive definite matrix and for 25 tests, 100% of our simulations resulted in a non positive definite matrix. We wanted to compare P_{ACT} with P_{rACT} but a direct application of P_{ACT} is not possible without some form of regularization since the covariance estimation may not result in a positive definite matrix otherwise. For P_{rACT} we regularized the matrix 4.9 using the method of Ledoit and Wolf [43]. In simulations our P_{rACT} was conservative; FWER based on $\alpha = 0.05$ was 0.0278 and 0.030 for 10 and 25 tests respectively.

False Discovery Rate adjustment for correlated tests

We would like to account for correlations between tests, as is done in the P_{ACT} FWER adjustment procedure, in the False Discovery Rate (FDR) framework of multiplicity adjustment. FDR adjustment controls the expected value of the unobservable random quotient Q of the number of true null hypotheses rejected V to the total number of hypotheses rejected R where

$$Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

Benjamini and Hochberg [3].

Benjamini and Hochberg [3] proposed a step-up procedure (BH) which controls the FDR for m tests at a specified level α. The BH procedure proceeds as follows. Let $p_{(1)} \leq$

$p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered unadjusted p-values; similarly $H^0_{(1)}, H^0_{(2)}, \ldots, H^0_{(m)}$ are the null hypotheses for the m tests in the corresponding order. Find the largest k such that $\frac{m \cdot p_{(k)}}{k} \leq \alpha$. If k exists, then reject the first k hypotheses $H^0_{(1)}, H^0_{(2)}, \ldots, H^0_{(k)}$ and if k does not exist then none of the hypotheses are rejected. The BH procedure is valid when the m tests are independent and some forms of dependence [4]. Our concern is that BH is overly conservative when the tests are positively correlated.

Here we propose an FDR adjustment procedure that accounts for correlations between the tests. We start by ordering the p-values as in BH. Then we apply the P_{ACT} procedure sequentially to each of the ordered p-values as described in Conneely and Boehnke [10]. If Σ is the estimated covariance matrix between the test statistics and the columns and rows are ordered according to the ordered p-values, we then compute the k-th adjusted p-value, $P_{(k)}$, using the $L \times L$ covariance matrix consisting of the k-th to m-th rows and columns of Σ with $L = m - k + 1$ and $p_{min} = p_{(k)}$ as described in [10]. In the BH procedure, we replace the expression $\frac{m \cdot p_{(k)}}{k}$ with $P_{(k)}$.

We test this modified BH procedure, BH_{ACT}, by simulating m correlated tests with $m_0 \leq m$ true nulls in a two sample problem where the outcomes for each observation are simulated using a multivariate normal distribution with an exchangeable correlation structure parameterized by ρ. We compute the observed false discovery rate across B simulations given α as

$$\widehat{\text{FDR}} = \frac{1}{B} \sum_{i=1}^{B} \frac{V_i(\alpha)}{R_i(\alpha)},$$

where $V_i(\alpha)$ is the number of rejected true null hypotheses and $R_i(\alpha)$ is the total number of rejected hypotheses at a level α for the i-th simulation. When exploring alternatives we compute the observed true discovery rate across the B simulations given α as

$$\widehat{\text{TDR}} = \frac{1}{B} \sum_{i=1}^{B} \frac{S_i(\alpha)}{m - m_0},$$

where $S_i(\alpha)$ is the number of rejected true alternative hypotheses at a level α for the i-th simulation.

We first simulate complete null data for ρ equal to 0, 0.1, 0.5 and 0.8 for ten tests with 50 observations evenly divided between two groups and run 1,000 simulations. Figure 4.4

shows that BH_{ACT} controls the FDR at the nominal rate for all simulated correlations while the standard BH procedure is progressively more conservative as the correlation increases.

Next we simulated tests where two of the ten outcomes differed between the two groups. For this analysis we considered ρ equal to 0 and 0.8 to test both the uncorrelated and highest correlated scenarios considered in the complete null case. For each value of ρ we considered four alternatives of varying magnitudes. On the standard deviation scale the magnitudes considered were 0.01, 0.1, 0.5, and 1 representing the mean difference between groups for the two true alternative hypotheses. Figure 4.5 shows that BH_{ACT} is more powerful than BH under correlated tests but slightly less powerful when the tests are uncorrelated. For correlated tests the gain in power of the ACT modification of the BH procedure is most noticeable when the magnitude of the alternative is moderate (e.g. 0.5 on the standard deviation scale).

4.8 Discussion

In this chapter we have defined a testing framework for K-mer scanning based on GEE and MO. We are interested in a method that remains unbiased when cluster size is informative therefore MO is the clear favorite over GEE in this respect. However, even if cluster size is informative, it is not clear that GEE is excessively biased under data simulations similar to the data we see in the Step trial. In order to see significant bias in the estimated treatment effect, we needed to simulate data with a very low level of correlation between sequences within a cluster. When a correlation closer to what we see in the real data was used we see very little bias in the estimated treatment effect (data not shown). MO on the other hand is very computationally expensive so we are interested in finding methods that will not require Monte Carlo simulation. In K-mer scanning we have the potential for many levels of Monte Carlo simulation. On top of the fact that we will have many tests to run when scanning the HIV-1 genome, we have potentially three separate levels of Monte Carlo simulation; 1) for a given K-mer each test might employ a permutation test as in the methods described in GWJ, 2) using MO to account for multiple sequences per subject, and 3) the P-value adjustment procedures of van der Laan et al. requires either bootstrap or permutation sampling. Given that we will need to run simulation studies to

fully understand the operating characteristics of any given method we have a total of up to five multiplicative factors all on the order of 10^3 that would necessitate on the order of 10^{15} operations when running a simulation study. For this reason we can only realistically expect to have one Monte Carlo step included among the three steps described above. For that reason we hope to find methods that can be performed analytically. Therefore, GEE methods are of great interest as they provide analytic methods which eliminate the need for Monte Carlo methods in steps 1 and 2.

CWGEE is asymptotically equivalent to MO so will likely remove the need to use MO to handle the scenario of informative cluster size. Another approach would be to consider the "true" cluster size to be the maximum observed cluster size and use weighted estimating equations as proposed by Robins, Rotnitzky, and Zhao (RRZ) [69] to account for missing sequence data as pointed out by Williamson et al. [84]. This would seem to be a reasonable interpretation for sequence data. In addition, the RRZ weighted estimating equations may be useful for missing data due to timing of sample collection. The most interesting K-mer scanning time point would be pre-sero conversion so that antibody pressure is not a factor in post-infection immune pressure. In this case all HIV-1 genomes sequenced post-sero conversion would be treated as missing.

As previously mentioned, interpretation of K-mer scanning, or any other high dimensional sieve analysis, requires methods for handling multiplicity adjustment. Although not commonly used there are multiplicity adjustment methods that incorporate the correlation structure between tests. For K-mer scanning, the tests are likely to be highly correlated so it seems that pursuing multiplicity adjustment methods that account for correlation, and potentially leveraging what is likely some type of auto-correlation structure, will likely provide additional power to our analyses.

In this chapter we have begun to incorporate the P_{ACT} method of multiplicity adjustment into the K-mer scanning framework but integration is incomplete. Clearly regularization of the sample covariance is needed to handle this high dimensional problem however there is more work needed to apply P_{ACT} when bias corrected GEE is used to generate the test statistics. Alternatively P_{ACT} when using multivariate MO is good candidate to pursue but the properties of FWER under this scenario need to be better characterized. Further-

more, although P_{ACT} has been touted by the authors of the method as computationally efficient yet it still requires the estimation of quantiles of a multivariate normal distribution which can be computationally time consuming and when combined with the computation time necessary to run MO the computing power necessary may become out of reach.

We have also explored FDR adjustment under a modification of the BH procedure using the ideas of P_{ACT}. Simulations suggest that this is a valid approach but proofs of control of FDR under independence and/or a complete null hypothesis are needed. Furthermore, application of the BH_{ACT} method would need to be tailored to the bias corrected GEE and MO methods to be of use to the K-mer scanning methodology presented in this chapter.

4.9 KC and mFG small sample adjustments are equivalent

To show this we first simplify the notation for the OPG adjustments used in the two procedures. Throughout this section I is an identity matrix with dimension implied by the context in which it is used. We want to show that the KC adjustment factor

$$\hat{D}_i^T \hat{\Sigma}_i^{-1} (I - \hat{H}_{ii})^{-1/2} = \hat{D}_i^T \hat{\Sigma}_i^{-1} (I - \hat{D}_i \hat{A}^{-1} \hat{D}_i^T \hat{\Sigma}_i^{-1})^{-1/2}$$

is equivalent to the mFG adjustment factor

$$\tilde{H}_i \hat{D}_i^T \hat{\Sigma}_i^{-1} = (I - \hat{D}_i^T \hat{\Sigma}_i^{-1} \hat{D}_i \hat{A}^{-1})^{-1/2} \hat{D}_i^T \hat{\Sigma}_i^{-1}.$$

To simplify notation we define

$$\begin{aligned} X &= \hat{D}_i^T \hat{\Sigma}_i^{-1} \\ Y &= \hat{D}_i \\ Z &= \hat{A}^{-1} \\ W &= (I - YZX)^{-1} \\ V &= (I - XYZ)^{-1}. \end{aligned}$$

In simplified notation we would like to show that $XW^{1/2} = V^{1/2}X$.

First we show that $XW = VX$ using the Sherman-Morrison-Woodbury identity.

$$\begin{aligned} XW = X(I - YZX)^{-1} &= X \left[I + Y(Z^{-1} - XY)^{-1} X \right] \\ &= \left[I + XY(Z^{-1} - XY)^{-1} \right] X \\ &= (I - XYZ)^{-1} X \\ &= VX \end{aligned}$$

then by induction we can show that

$$X(W - I)^n = (V - I)^n X$$

for all integers n. The identity above proves the case for $n = 1$ and given $X(W - I)^k =$

$(V-I)^k X$,

$$\begin{aligned} X(W-I)^{k+1} &= X(W-I)^k(W-I) \\ &= (V-I)^k X(W-I) \\ &= (V-I)^k(V-I)X \\ &= (V-I)^{k+1}X \end{aligned}$$

which proves the induction step.

Finally, using a taylor expansion of $W^{1/2}$ we have

$$\begin{aligned} XW^{1/2} &= X\left[I + \frac{1}{2}(W-I) + \sum_{n=2}^{\infty}(-1)^{n+1}\frac{(2n-3)!}{2^{n-2}n!(n-2)!}(W-I)^n\right] \\ &= \left[I + \frac{1}{2}(V-I) + \sum_{n=2}^{\infty}(-1)^{n+1}\frac{(2n-3)!}{2^{n-2}n!(n-2)!}(V-I)^n\right]X \\ &= V^{1/2}X. \end{aligned}$$

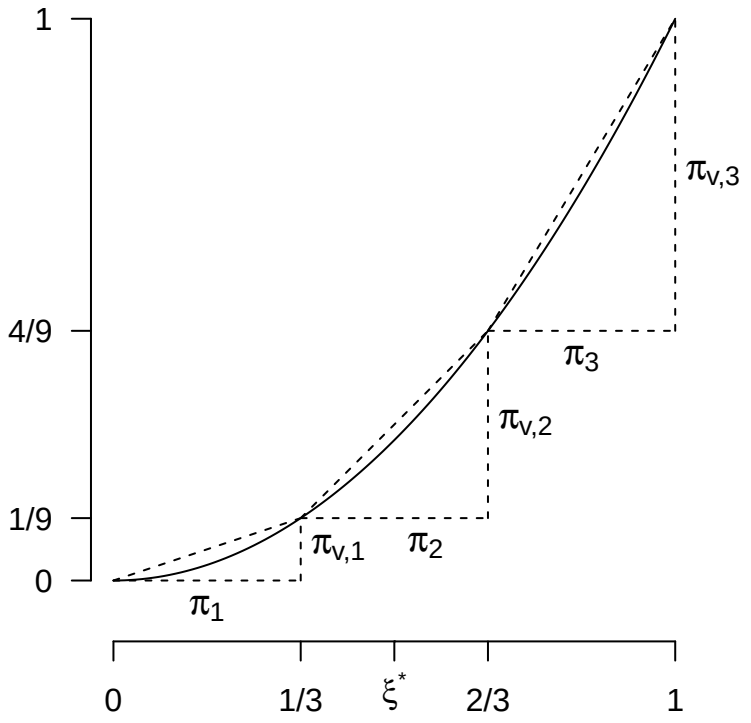

Figure 4.1: An example of cut point and probability transformations, when $\epsilon = 2$, used to generate alternative hypotheses. The solid line shows the function $f(x) = x^2$ used to transform the cut points, $\boldsymbol{\xi} = (1/3, 2/3, 1)$, into the cut points, $\boldsymbol{\xi}_v = (1/9, 4/9, 1)$. The corresponding probabilities of the vectors $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_v$ are shown on each leg of the three dashed triangles. The slope of each secant gives the scale factors which transform the probability vector, $\boldsymbol{\pi}$, to the probability vector, $\boldsymbol{\pi}_v$. The x-value, ξ^*, is the point at which the derivative of f is equal to one which marks the cut point transition between scale factors which are less than one and those which are greater than one.

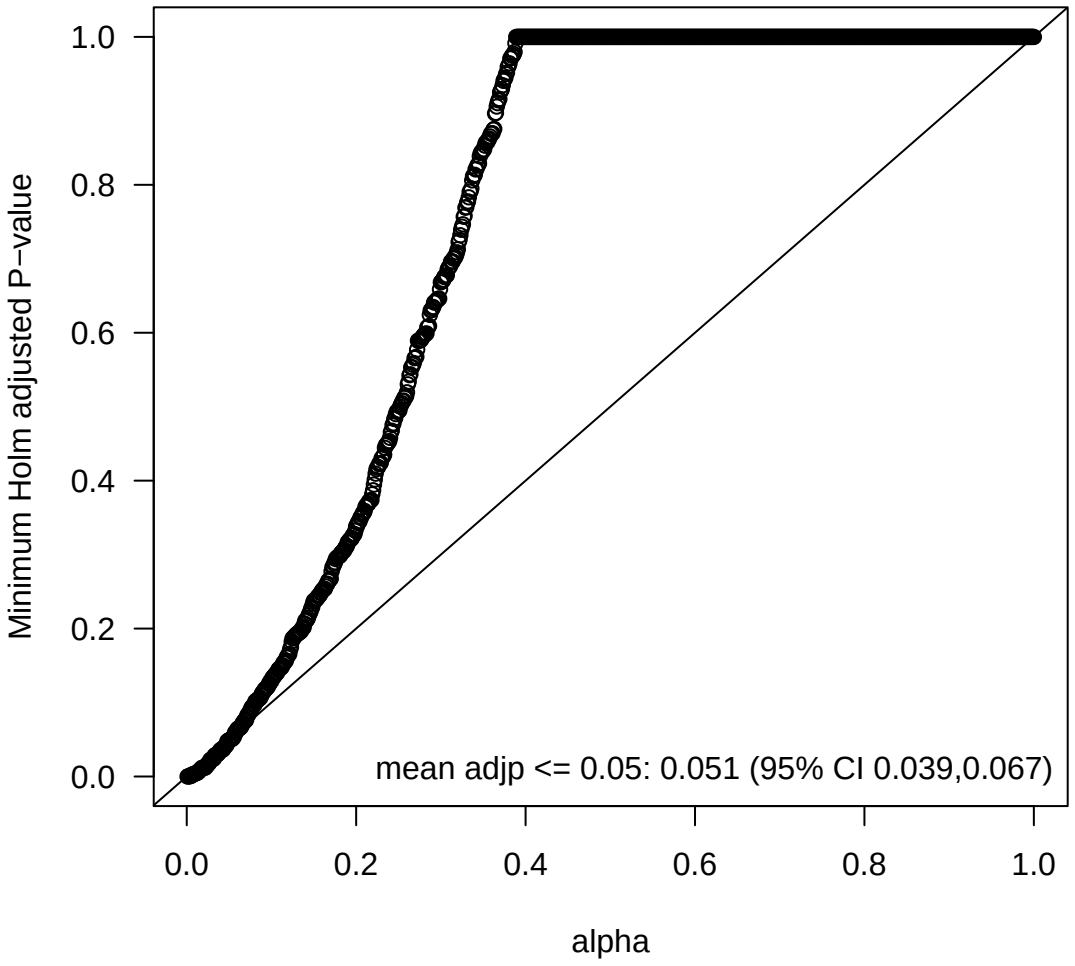

Figure 4.2: Probability-probability plot assessing FWER. Using Method 2, 1,000 simulations of the complete null hypothesis were generated with $\rho_1 = 0.9$and $\rho_2 = 0.5$ using Gag breakthrough sequences from the Step trial placebo group. The 9-mer scanning method using GEE was applied to the 534 positions (542 alignment positions plus length of a 9-mer minus 1). For each simulation we compute the minimum Holm adjusted P-value. For each adjusted P-value we plot the proportion of adjusted P-values that are less than or equal to that value. The identity line is plotted for reference.

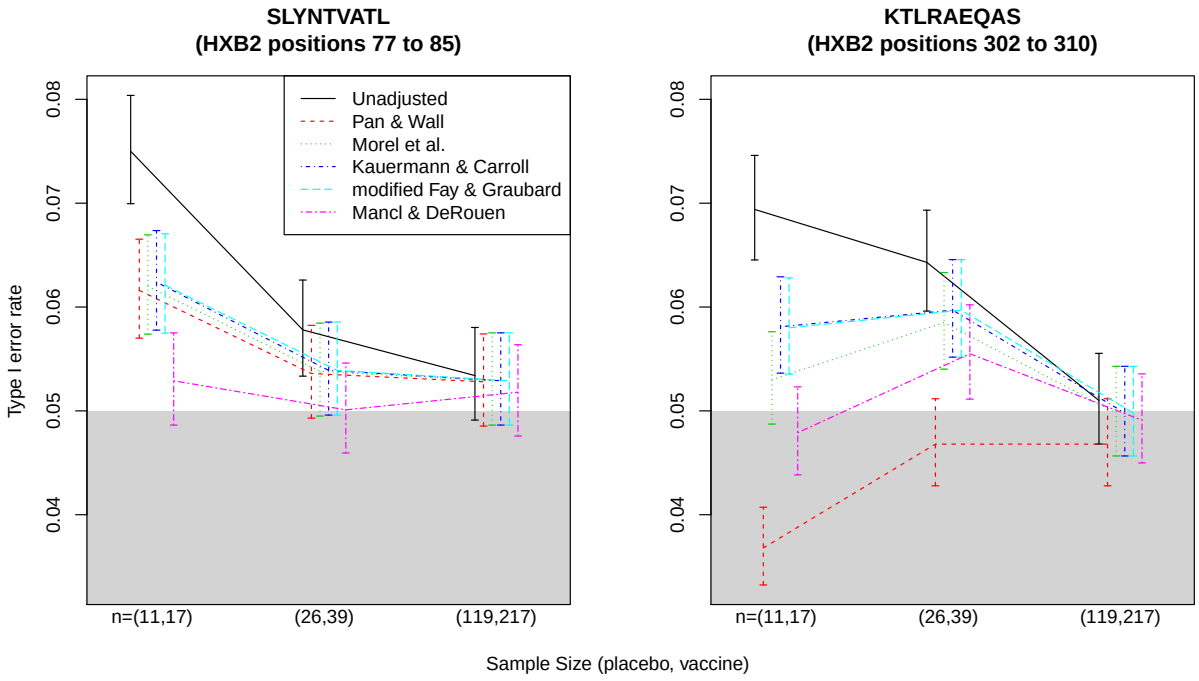

Figure 4.3: Sandwich Estimator Bias Corrections. Point estimates and 95% confidence intervals of the level 0.05 Type I error rate are shown for the six variance estimators listed in the left side panel. The estimates were generated based on 10,000 simulations at two K-mer regions SLYNTVATL and KTLRAEQAS. Three sample sizes noted on the x-axis were simulated. The first is based on the number of A02 positive participants from the Step sieve cohort with 11 placebo and 17 vaccine recipients. The second is the entire Step sieve cohort with 26 placebo and 39 vaccine recipients. The third sample size is based on the Vax004 sieve cohort with 119 placebo and 217 vaccine recipients. The gray band at the bottom of the figure marks the transition between conservative and liberal Type I error control for $\alpha = 0.05$.

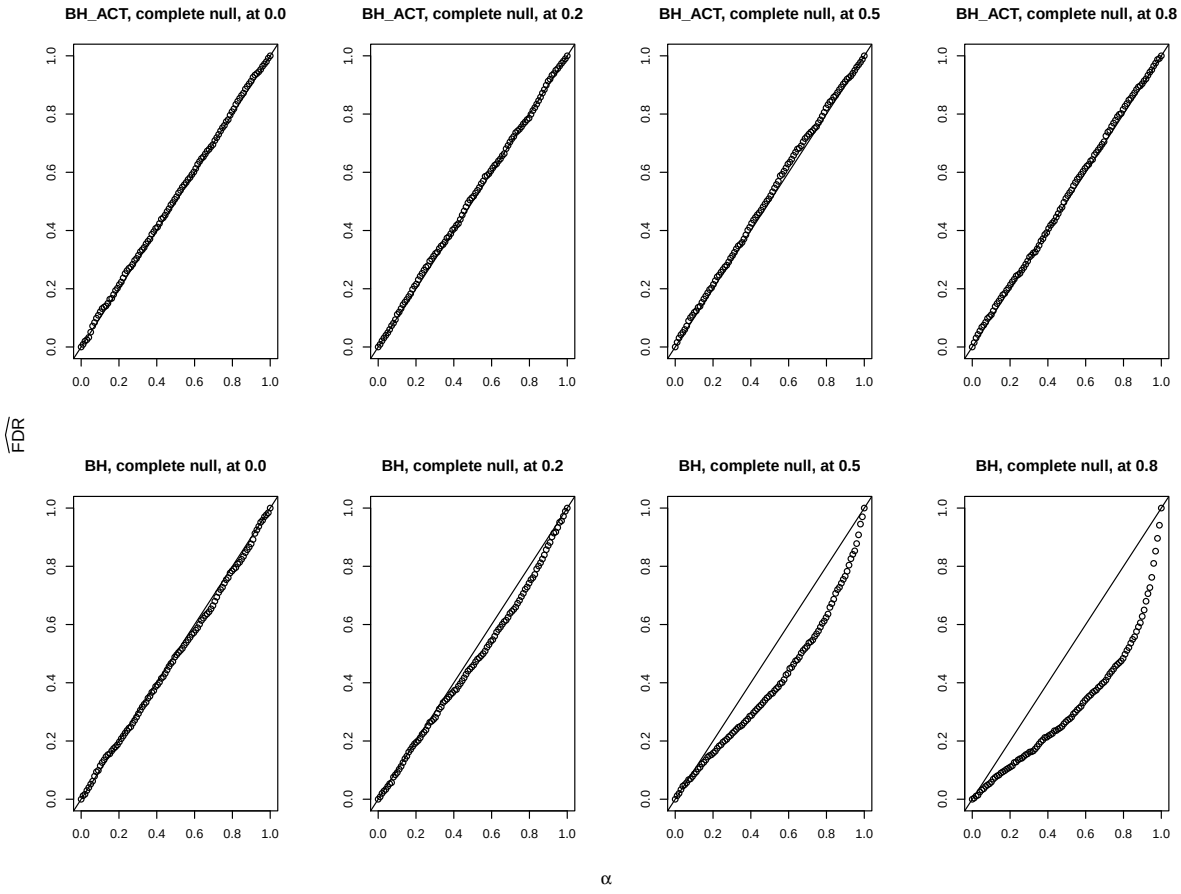

Figure 4.4: Comparison of the observed versus nominal FDR for 1,000 simulations of the complete null for 10 tests with exchangeable correlation between tests. The top row shows the results for the BH_{ACT} procedure and the bottom row shows the results for the BH procedure. For each simulation, 50 observations each consisting of 10 correlated outcomes were simulated using a multivariate normal distribution with exchangeable correlation of 0, 0.1, 0.5 or 0.8. The solid line indicates FDR equal to α.

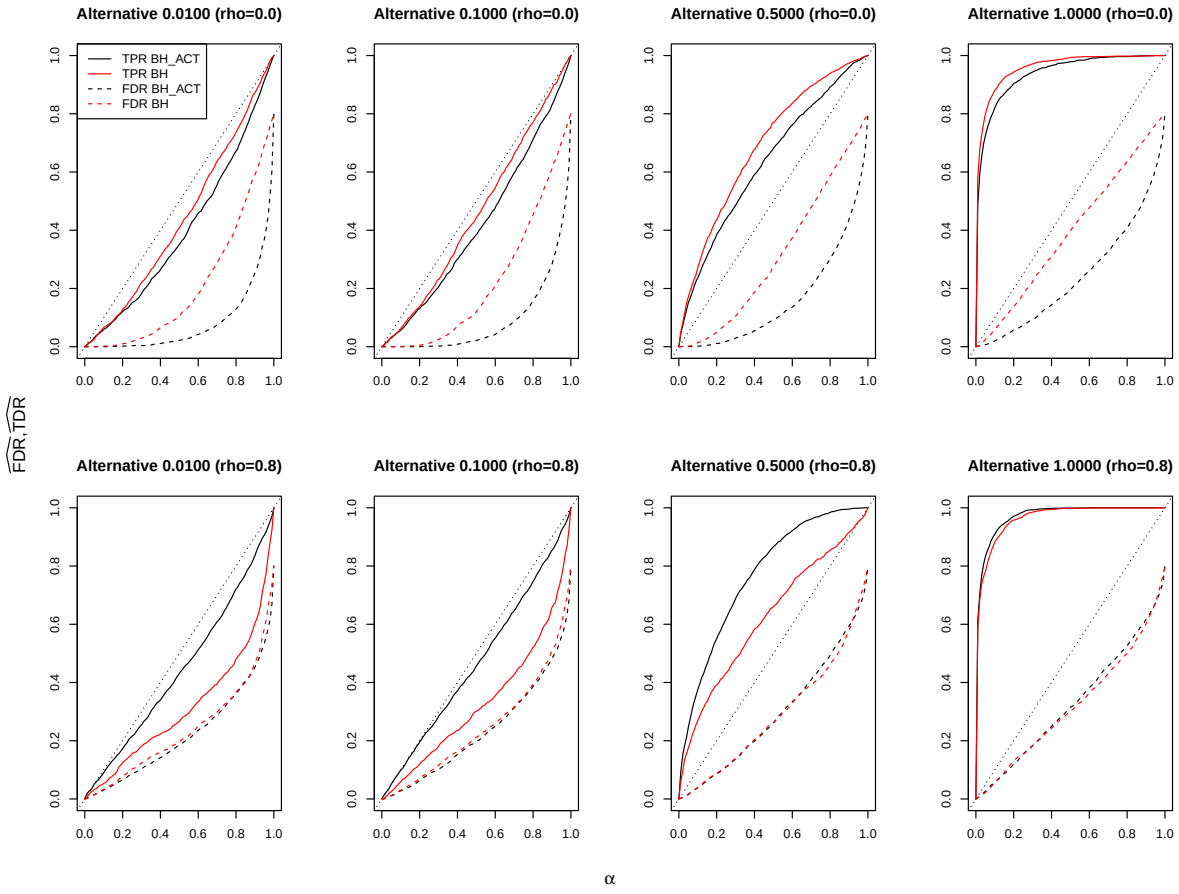

Figure 4.5: Comparison of the observed versus nominal FDR and TDR for 1,000 simulations of 10 tests with exchangeable correlation between tests with two true alternative hypotheses and eight true null hypotheses. The top row shows the results for uncorrelated tests while the bottom row shows the results for tests with high correlation. The magnitudes of the effect on the standard deviation scale for the difference in group means is 0.01, 0.1, 0.5, and 0.8. For each simulation, 50 observations each consisting of 10 correlated outcomes were simulated using a multivariate normal distribution with exchangeable correlation of 0, 0.1, 0.5 or 0.8 and means equal to zero except for the two true alternative hypotheses in which case the outcomes for one group had a mean given by the magnitude. The vertical axis shows the FDR/TDR while the horizontal axis shows the nominal level α. The BH_{ACT} procedure results are shown in red (FDR is the dashed line and TDR is the solid line) while the the results for the BH procedure are shown in black (again, FDR is the dashed line and TDR is the solid line).

K-mer	HXB2	Test	Size (95% CI)
SLYNTVATL	77:85	MO	0.051 (0.047,0.056)
SLYNTVATL	77:85	GEE	0.064 (0.059,0.069)
SLYNTVATL	77:85	CWGEE	0.064 (0.059,0.069)
KNCRAPRKK	403:411	MO	0.053 (0.049,0.058)
KNCRAPRKK	403:411	GEE	0.061 (0.057,0.066)
KNCRAPRKK	403:411	CWGEE	0.060 (0.056,0.065)

Table 4.1: Size is the observed Type I error rate for $\alpha = 0.05$ in 10,000 simulated Null distributions using simulation Method 1. A 95% confidence interval for size is included. K-mer is the vaccine strain peptide at the specified HXB2 location in Gag used when generating the simulation data.

K-mer	HXB2	Beta	Method	Mean	Power (95% CI)
SLYNTVATL	77:85	-17.14	MO	-17.19	0.860 (0.853,0.867)
SLYNTVATL	77:85	-17.14	GEE	-17.12	0.857 (0.850,0.864)
SLYNTVATL	77:85	-17.14	CWGEE	-17.12	0.856 (0.849,0.863)
KNCRAPRKK	403:411	-15.16	MO	-15.23	0.864 (0.857,0.870)
KNCRAPRKK	403:411	-15.16	GEE	-15.15	0.862 (0.855,0.868)
KNCRAPRKK	403:411	-15.16	CWGEE	-15.14	0.861 (0.854,0.867)

Table 4.2: Power is the observed proportion of rejected null hypotheses for $\alpha = 0.05$ in 10,000 simulations of the Method 1 non-informative cluster size alternative for $\epsilon = 2.5$ and $\rho = 0.9$. For MO, resampling was done 100 times. Beta is the true treatment effect; values are negative since under the alternative K-mers are less similar than the vaccine insert in the vaccine group than the placebo group. Mean is the mean of the estimated Beta values. A 95% confidence interval for power is included. Simulated data are based on the location indicated by K-mer which is the vaccine strain peptide at the specified HXB2 location in Gag.

K-mer	HXB2	Beta	Method	Mean(Beta)	Coverage (95% CI)
SLYNTVATL	77:85	-17.14	MO	-17.19	0.944 (0.939,0.948)
SLYNTVATL	77:85	-17.14	GEE	-17.12	0.935 (0.930,0.939)
SLYNTVATL	77:85	-17.14	CWGEE	-17.12	0.936 (0.931,0.940)
KNCRAPRKK	403:411	-15.16	MO	-15.23	0.940 (0.935,0.944)
KNCRAPRKK	403:411	-15.16	GEE	-15.15	0.940 (0.935,0.944)
KNCRAPRKK	403:411	-15.16	CWGEE	-15.14	0.939 (0.934,0.944)

Table 4.3: Coverage is the observed proportion 95% confidence intervals for the estimated Beta that include the true value of Beta in 10,000 simulations of the Method 1 non-informative cluster size alternative for $\epsilon = 2.5$ and $\rho = 0.9$. For MO, resampling was done 100 times. Beta is the true treatment effect; values are negative since under the alternative K-mers are less similar than the vaccine insert in the vaccine group than the placebo group. Mean is the mean of the estimated Beta values. A 95% confidence interval for coverage is included. Simulated data are based on the location indicated by K-mer which is the vaccine strain peptide at the specified HXB2 location in Gag.

K-mer	HXB2	Beta	Method	Mean	Power (95% CI)
SLYNTVATL	77:85	-10.18	MO	-10.34	0.887 (0.865,0.906)
SLYNTVATL	77:85	-10.18	GEE	-7.73	0.842 (0.818,0.864)
SLYNTVATL	77:85	-10.18	CWGEE	-10.06	0.865 (0.842,0.885)
KNCRAPRKK	403:411	-9.28	MO	-9.38	0.910 (0.890,0.927)
KNCRAPRKK	403:411	-9.28	GEE	-7.15	0.863 (0.840,0.883)
KNCRAPRKK	403:411	-9.28	CWGEE	-9.14	0.882 (0.860,0.901)

Table 4.4: Power is the observed proportion of rejected null hypotheses for $\alpha = 0.05$ in 100 simulations of the Method 1 informative cluster size alternative for $\epsilon = 2$, $\eta = 5$, and $\rho = 0.1$. For MO, resampling was done 1,000 times. Beta is the true treatment effect; values are negative since under the alternative K-mers are less similar than the vaccine insert in the vaccine group than the placebo group. Mean is the mean of the estimated Beta values. A 95% confidence interval for power is included. Simulated data are based on the location indicated by K-mer which is the vaccine strain peptide at the specified HXB2 location in Gag.

K-mer	HXB2	Beta	Method	Mean	Coverage (95% CI)
SLYNTVATL	77:85	-10.18	MO	-10.34	0.953 (0.937,0.965)
SLYNTVATL	77:85	-10.18	GEE	-7.73	0.816 (0.790,0.839)
SLYNTVATL	77:85	-10.18	CWGEE	-10.06	0.920 (0.901,0.936)
KNCRAPRKK	403:411	-9.28	MO	-9.38	0.951 (0.935,0.963)
KNCRAPRKK	403:411	-9.28	GEE	-7.15	0.818 (0.792,0.841)
KNCRAPRKK	403:411	-9.28	CWGEE	-9.14	0.940 (0.923,0.954)

Table 4.5: Coverage is the observed proportion 95% confidence intervals for the estimated Beta that include the truth value of Beta in 10,000 simulations of the Method 1 informative cluster size alternative for $\epsilon = 2$, $\eta = 5$, and $\rho = 0.1$. For MO, resampling was done 1,000 times. Beta is the true treatment effect; values are negative since under the alternative K-mers are less similar than the vaccine insert in the vaccine group than the placebo group. Mean is the mean of the estimated Beta values. A 95% confidence interval for coverage is included. Simulated data are based on the location indicated by K-mer which is the vaccine strain peptide at the specified HXB2 location in Gag.

Chapter 5

ASSESSING CYTOTOXIC T LYMPHOCYTE EFFECTS

5.1 Introduction

Major histocompatibility complex (MHC) binding assays measure the ability of MHC molecules to bind peptides. Viral peptides bound to MHC molecules form an immune complex which provides an immune surveillance system. T-cells that recognize these immune complexes on the surface of infected cells through T-cell receptors (TCRs) can directly kill the infected cell or generate signaling that prompts other immune system cells to respond. In this chapter we focus on methods for assessing cytotoxic T lymphocyte (CTL) effects. Specifically, CTL effects are mediated through the recognition of MHC Class-I molecules bound to short peptides, typically of length between 9 and 11 amino acids, by the TCRs of CD8 positive T-cells. In humans, MHC molecules are produced by HLA alleles and the MHC molecules are also referred to as HLA molecules.

MHC Class-I presentation of viral antigens by infected cells requires the following steps. First, intracellular viral proteins are degraded by the proteasome into peptides. These peptides are transported to the endoplasmic reticulum where they bind with varying affinity to MHC Class I molecules. Binding affinity depends on host MHC alleles and viral peptide amino acid composition. The immune complex consisting of the bound MHC molecule and peptide are then transported to the plasma membrane where they are visible to CD8 positive T-cells. If the MHC/peptide complex is recognized by a CD8 T-cell with a TCR specific to that MHC/peptide combination, the T-cell can destroy the infected cell. Although this is a very brief description of an extremely complex and nuanced immune response, the sequence described above is comprised of three basic components: 1) peptide cleavage, 2) peptide binding, and 3) TCR recognition. In this chapter, our analysis of CTL effects utilizes on the critical second step which requires the ability of an MHC molecule to bind a specific peptide within the vaccine construct. By adding host genetic information regarding HLA alleles with MHC binding prediction we hope to focus the K-mer scanning method of the previous chapter on subject specific targets of immune pressure improving the power of the previous methods as well as specifically addressing CTL induced immune pressure.

MHC binding assays measure the ability of peptides to bind to MHC molecules. These assay readouts often report the concentration of peptide that displaces 50% of a reference

peptide from the MHC molecule known as the IC_{50}. MHC binding prediction algorithms attempt to predict the level of binding of an MHC molecule to a peptide given only the MHC allele and the amino acid sequence of the peptide.

One well known and often study phenomenon observed in HIV infected individuals is CTL induced escape. Selective HLA-restricted mutations often occur in epitopes recognized by CD8 cells with a TCR specific for a given MHC/peptide combination allowing HIV to evade CTL immune surveillance. Mutations that prevent or lower the binding affinity between the host MHC molecules and viral peptides is one mechanism of CTL induced escape. The goal of the method described in this chapter is to assess differential escape patterns between treatment arms in a vaccine trial. We use MHC binding prediction algorithms and differences in the predicted binding levels across peptides within a gene between the vaccine strain and the breakthrough strains in an attempt to capture a count of potential CTL escapes or more generally to quantify the amount of CTL pressure. We refer to this as potential T-cell escape (PTE). Our hypothesis is that for some vaccine regimens CTL pressure, as measured by PTE, will differ between the treatment arms. We initially propose a method in which we cannot infer causal relationships. However, potential outcomes as define by Rubin [73] fit naturally into our proposed analysis and we expand the method to explore causal inference in Section 5.6.

In this chapter we present prior work in Section 5.2, we introduce our method by defining binding scores in Section 5.3, we demonstrate our method on simulated data in Section 5.4, we analyze the Step trial data in Section 5.5, we explore extending the method to a causal inference framework in Section 5.6, and end with a discussion in Section 5.7.

5.2 Prior work

When applied to contiguous K-mers using lengths between 9 and 11, the work of the previous chapter can be viewed as one method for assessing a vaccine induced CTL effect. This approach has been applied previously to breakthrough sequences from the Step trial [71] as well as the Phambili and RV144 trials (not yet published). The K-mer scanning method however ignores host genotype information which should be valuable for focusing the analysis on relevant host-specific targets of the immune system. As discussed above,

one of the critical steps in the CTL response is the binding of viral peptides to host HLA molecules. By incorporating MHC binding prediction algorithms we hope to focus the analysis on the most relevant targets. When we do this, the data become much more sparse for a given K-mer within the vaccine. The number of subjects with an HLA molecule that is predicted to bind varies widely but often a K-mer is recognized by only a small fraction of the participants. For this reason we intend to aggregate across K-mers only counting those that have predicted binding (or, if available, observed binding) at some pre-specified threshold or range of thresholds.

The percent epitope mismatch and predicted epitope distances from [71] are examples where the distance is computed over an entire gene represented by the vaccine construct and incorporate HLA data. Additionally, in Janes et al. [36], an aggregate measure across each gene was generated using observed T-cell responses within the vaccine group prior to infection rather than using MHC binding predictions. Using observed T-cell responses has the added benefit that the full pathway of CD8 T-cell recognition, rather than just the MHC/peptide binding step, must have been satisfied which further focuses the analysis on relevant sites of immune pressure. The downsides are 1) these measurements are only observable and meaningful in the vaccine group since placebo subjects have had no prior exposure to HIV-1 viral antigens; 2) the ability of in vitro assays to detect CD8 T-cell responses relevant to immune pressure may be limited; and 3) post-infection responses may not be accurately captured by pre-infection vaccine induced responses.

Here we expand the previous methods that use MHC binding algorithms to 1) account for differential binding between the vaccine reference strain and breakthrough sequences and 2) use the viral genetic variability in breakthrough viruses at regions of predicted non-MHC binding.

5.3 Binding Scores

To assess potential vaccine induced CTL effects we first fix a K-mer region within the vaccine insert and fix an HLA allele. For example, consider the Gag 9-mer region corresponding to the HXB2 positions 77-85. This region has been well studied and HIV-infected individuals who express HLA-A*02:01 typically recognize the SLYNTVATL epitope SL9 [77]. For each

infection we will have a breakthrough sequence and the subject's HLA genotype. These form a random variable E (for epitope) which is the joint distribution of P, the breakthrough peptide in the region of interest, and H, the subject's HLA type. For a given k-mer, p, and allele, h, we can use a binding prediction algorithm to assign a binding value $B_h(p)$. We drop the allele when the notation is unambiguous and without loss of generality we assume that a low binding value represents tight binding while a high value indicates poor binding which is the case when binding is measured as IC_{50}.

In a vaccine trial we will have a reference K-mer, r, in the region of interest and we define the binding score as a weighted difference in the binding values for P and r given an HLA allele h. When computing a binding score, we consider h rather than H since h need not be the same HLA allele as the subject's HLA type H encoded in the epitope variable E. Define,

$$S(P, r, h) = w \times (B_h(P) - B_h(r))$$

where $w \in [0, 1]$. This score is designed to reflect an *immune pressure model.* In this model, we want the binding score to be positive when allele h is equal to H and h allele exerts pressure on the virus away from the reference k-mer. Otherwise, we want the score to be close to zero. A slightly more general form of the binding score is

$$S(P, r, h) = w \times f\left(B_h(P) - B_h(r)\right) \tag{5.1}$$

where f is an increasing function. As discussed in the introduction to this chapter, we may want to capture PTE as a count of potential CTL escape mutations for each subject in which case $f : \mathbb{R} \to \{0, 1\}$ or more generally we may want to assess the probability of an escape in which case $f : \mathbb{R} \to [0, 1]$.

In general, the mean and variance of $S(P, r, h)$ will depend on the amount of agreement between circulating strains and the ability of allele h to apply immune pressure on the reference peptide. We now consider specific definitions of w and f in equation 5.1 in an attempt to reflect our understanding of the biology of HLA-associated escape in the *immune pressure model.*

First the weight w accounts for reference K-mers that are poor binders. We may see a large difference in binding between P and r but if the binding of r to the MHC molecule

produced by h is weak, we still want $S(P, r, h)$ to be close or equal to zero. A simple weight, the *binding threshold weight*, is an indicator based on a binding threshold γ,

$$w_\gamma = \mathbf{I}_{[B_h(r) \leq \gamma]}.$$

A slight variation on this would be to use a probit function of $B_h(r)$. We are not interested in the parameter γ (nor probit function parameters) as it is a biological constant determined by the specific method for assigning binding values. For example, we might use γ values between the strong (50 nM) and weak (500 nM) binder definitions defined by the netMHC binding predictor[55].

Next we account for negative and small positive differences in binding values. These scores are unlikely to be the result of immune pressure and should be down weighted or set to zero. The *binding delta function* is defined as $f_\delta(P, r, h) = \mathbf{I}_{[B_h(P) - B_h(r) \geq \delta]}$. Delta is a nuisance parameter; we could either estimate δ based on past experience, use a range of δs, or integrate over all possible δs.

Combining the *binding threshold weight* with the *binding delta function* we have one form of the binding score function

$$S(P, r, h) = w_\gamma(r, h) \times f_\delta(P, r, h) = \mathbf{I}_{[B(r) \leq \gamma]} \times \mathbf{I}_{[B_h(P) - B_h(r) > \delta]}. \tag{5.2}$$

5.3.1 Testing

A simple global test for treatment effect would first compute an average S across all K-mers and HLA alleles for each subject then compare average binding score across treatment arms using a t test or GEE if multiple sequences per subject were used.

5.3.2 Alternative to using a "delta binding function"

From 93- to 99-percent of K-mers will be non-binders[1] and we ensure that most of the scores based on these K-mers will be assigned a binding score close to zero by including the *binding threshold weight*. It's only for an HLA-restricted epitope that we expect immune pressure

[1]Based on the ADT predicted binding of all majority consensus sequence 9-mers and 4-digit HLA-A and HLA-B alleles from the 64 breakthrough infections used in the Step Gag analysis at levels of nM 500 and nM 50, respectively.

and a high binding score. Our difficulty lies in distinguishing between high binding scores due to immune pressure and high binding scores due to mismatches between circulating strains and the reference strain. In the framework described above we ignored a potentially valuable source of information in the binding scores for subjects with mismatched alleles. Even if a subject does not have allele h we can still compute the binding score $S(P, r, h)$ using this allele and these binding scores across subjects contain information about the variability in the levels of binding due to sequence variability in the region of interest.

We expect to gain power for detecting a treatment effect by including the binding scores of the subjects with a mismatched HLA allele because their sequence information will be useful in computing the "null" distribution of binding scores in the absence of pressure. Potentially a large percentage of subjects place pressure on some regions (e.g. hotspot regions). This might cause problems with the assumption that $S(P, r, h)$ is positive when r is under pressure for $H = h$ and close to zero for $H \neq h$. One mitigating factor is that pressure from differing alleles may not have the same escape pattern(s). Therefore we might expect $S(P, r, h)$ to be positive even if pressure is applied by multiple HLA alleles to the same region. HLA alleles are sparse in most vaccine trial populations in that for any given allele typically only a small fraction of the study population will have that allele. This is actually beneficial in this framework. Finally, a potential method for mitigating problems with pressure from other alleles would be to dichotomize the alleles as h and those not related to h. This could be done using HLA supertypes or when using 4-digit typing by only considering subjects with an HLA genotype that does not match h at the 2-digit level. This would eliminate some potentially valuable information but would probably remove any problems associated with multiple alleles that produce the same escape pattern(s).

We implement this idea by modifying the *delta binding function* of equation 5.2 using an alternative threshold function calculated using data available for $H \neq h$. First, consider a vaccine efficacy trial with n subjects. Now, fix a peptide r within the vaccine. For each infected subject assume we identify a single breakthrough virus and align that virus to the vaccine reference strain. For each infected participant $i = 1 \ldots n$, we have a peptide P_i and

a set of HLA alleles $\mathcal{H}_i$. Define a threshold function

$$\tau(r, h) = \text{median}\{B_h(P_i)|h \notin \mathcal{H}_i\}. \tag{5.3}$$

Our alternative to using the "delta binding function" is to define

$$S(P, r, h) = \mathbf{I}_{[B_h(r)\leq\gamma]} \times \mathbf{I}_{[B_h(P)-B_h(r)>0]} \times \mathbf{I}_{[B_h(P)>\tau(r,h)]}. \tag{5.4}$$

Notice that we have replaced δ in equation 5.2 with zero and added an indicator function that requires the binding of P to be greater than the threshold defined in equation 5.3. We consider the triple (P, r, h) a potential escape if three conditions are satisfied: 1) r binds with high affinity to the HLA molecule h, 2) peptide r binds with higher affinity (i.e., lower IC_{50}) to h than peptide P, and 3) peptide P binds with lower affinity than the median predicted binding level, given h, of observed peptides aligned to r that are not under immune pressure.

5.3.3 Testing

Testing can be done by an appropriate two sample testing method. Let $i = 1, \ldots, n$ index the breakthrough infections in a trial and let $\mathcal{R}$ be the set of reference peptides within the vaccine strain. For each subject i, we have a set of peptides $\{P_{i,r}|r \in \mathcal{R}\}$ indexed by $\mathcal{R}$ that have been aligned to the reference peptides. For each i, we also have the set of unique 4-digit HLA alleles $\mathcal{H}_i$; we only take one allele given a homozygous pair. Then for each subject i we define the PTE outcome as

$$Y_i = \sum_{r\in\mathcal{R}} \sum_{h\in\mathcal{H}_i} S(P_{i,r}, r, h). \tag{5.5}$$

For each subject we have and indicator of treatment assignment Z_i where 0 indicates placebo and 1 indicates active treatment. We test the null hypothesis

$$H_0 : E[Y|Z = 0] = E[Y|Z = 1]$$

that the expected value of Y is equal in the two treatment groups.

5.4 Data simulation

Consider placebo controlled randomized HIV-1 vaccine efficacy trial. Even though a vaccine may be ineffective at blocking infection it may have a positive effect on disease progression via accelerated T-cell mediated immune pressure. We would like to simulate data from an efficacy trial in which we collect HLA class I genotype information from all subjects and a single acute (i.e., pre-ELISA positive) sequence from each subject.

Brumme et al [7] analyzed greater than 1,500 mostly subtype B sequences for associations between HLA allele and amino acid frequency. Depending on q-value threshold they identified between 80 (of 500; 16%) and 130 (26%) codons in Gag with HLA associated polymorphisms. Using the Step study breakthrough sequences and the polymorphisms identified by Brumme et al we simulate a data set consisting of acute sequences and subject HLA genotypes designed to emulate a vaccine that accelerates HLA mediated immune pressure.

We simulate data with the same number of placebo ($n = 25$) and vaccine ($n = 39$) Gag sequences used in the study by Rolland et al. [71]. We start by sampling 64 of the 65 Step breakthrough consensus Gag sequences (one subject is missing HLA data which is why Rolland et al. only used 64 sequences). Next we sample a set of HLA-A and -B alleles by independently sampling the 4 HLA-A and HLA-B alleles from the 64 observed HLA genotypes. Although this does not account for potential HLA linkage disequilibrium it does reasonably simulate HLA data for the infected participants in the Step trial. This HLA sampling also breaks any association between the observed HLA genotypes and breakthrough sequences. Finally, we randomly assign the set of treatment assignments to these 64 sequence/HLA genotype pairs with the same frequency observed in the Rolland study. In this way we generate a set of acute infection sequences along with a set of HLA genotypes simulating the data from an HIV-1 vaccine efficacy trial. At this point the sequences are the observed sequences for the Step trial, however, any association between sequences and HLA and/or treatment should be broken by the random generation of HLA genotypes and random treatment assignment.

The next step is to apply HLA specific mutations in the vaccine group to simulate an

accelerated immune response. We do not simulate immune escape in the placebo group since we assume that the breakthrough sequences are discovered early enough to have limited or no escape due to T-cell pressure. Our goal is to simulate three levels of escape; low, medium and high. We begin by briefly reviewing the types of associations found by Brumme et al. Amino acids which are significantly enriched in the presence of an HLA allele are referred to as adapted to that allele. If the amino acid is significantly depleted it is referred to as nonadapted. Associations are also classified as direct, mutations that are a direct result of CTL pressure, or indirect, associations due to compensatory or secondary mutations. Since adapted associations with an amino acid that matches the vaccine strain would not represent pressure due to the vaccine we do not use these associations in our simulations. Each association specifies a codon, a 2-digit HLA type, and an amino acid. For each adapted association we consider all sequences from subjects with a matching HLA allele in our resampled data set and mutate those sequences that do not already match the adapted amino acid. For each nonadapted association we again consider the sequences of subjects with a matching HLA allele but this time we redistribute the nonadapted amino acid among the amino acids observed at this codon that differ from the vaccine strain amino acid.

The binding escape score in equation 5.4 is only defined for peptides P that align to the reference peptide r. There are several ways one could handle this alignment issue but we take a simple approach in our simulation study. Starting with the 65 consensus Gag sequences from Step multiply aligned with the vaccine strain, we select the 303 9-mer peptides r that have a complete set of matching 9-mers among the breakthrough consensus sequences. That is we only use the 303 positions that have contiguous 9-mers, without insertions and deletions, among the 65 aligned sequences. When simulating mutations using the Brumme associations we only accept a mutation if all the resulting 9-mer peptides in the simulated sequence are among the observed 9-mer peptides across the selected positions. The reasoning for this restriction is two-fold. The first reason is a purely practical consideration; we do not need to compute additional binding predictions beyond those used to run the method on the original data. The second reason is this requirement means that simulated sequences will maintain the characteristics of real sequences at least at the level of 9-mer peptides, which is the main unit of analysis; that is although the simulated Gag sequences in the vaccine

group are not likely to match one of the observed Gag sequences every 9-mer derived from those sequences and used to compute the binding score will be an observed peptide.

The simulation process described to this point defines the lowest level of escape. It may at first seem that there is no way to create higher levels of escape within this framework because we use every potential association to generate escapes. The way we generate higher levels of escape is to define multiple HLA genotypes for each sequence in the vaccine group and pick the HLA genotype that generates the highest number of mutations. Two HLA genotypes per simulated vaccine recipient sequence were used to simulate medium escape and three HLA genotypes per sequence were used to simulate the high level of escape.

We simulated 1,000 data sets of 25 placebo sequences and HLA genotypes along with 39 vaccine sequences and genotypes for the three levels of escape. We also generate 1,000 null data sets by sampling HLA genotypes and sequences as described above without applying any simulated mutations.

5.4.1 Data simulation results

The frequency of HLA alleles in the simulated data sets is defined by the observed frequencies of 64 infected trial participants from the Step study. The distribution of HLA alleles in the simulated placebo recipients maintain these frequencies by design of the study. This distribution of HLA alleles is also maintained in the vaccine recipients for the null data sets and the "Low" level of mutation. The vaccine group has a different distribution of HLA alleles for the "Medium" and "High" levels of mutation because some alleles are more likely to produce an accepted mutation within the simulation study design. Figure 5.1 shows the enrichment and depletion of the HLA allele frequencies for "Medium" and "High" mutation scenarios compared to the base frequencies. For example, HLA-B*07 is enriched under both scenarios while HLA-B*18 is depleted. Notice that, the directionality of the frequency changes tend to be the same across alleles for both of these scenarios while the magnitude of the change tends to be highest for the "High" mutation scenario. Across the three mutation scenarios the number of simulated participants with homozygous alleles in the vaccine group tends to decrease as the mutation rate increases (see Table 5.1).

The average number of mutations per sequence and the average number of mutated sites per data simulation for the vaccine group sequences are shown in Table 5.2. As expected, the number of mutations per sequence and the number of sites with one or more mutated sequences within a simulated data set increase as the level of mutation increases. The mean number of mutations per sequence are low with only 2.77 mutations per vaccine sequence in the "High" level of mutation. These mutations are distributed over a fairly broad number of sites with on average almost 30 sites having one or more sequences with a simulated mutation at that site. These characterizations of the data simulations show that we have generated data that achieves a low frequency of simulated escape that is distributed across a range of sites. Only a method that allows a subject specific targeting of the true escape locations will have any power to detect this type of alternative.

We use the testing method described in Section 5.3.3 and use the ADT prediction algorithm to compute predicted binding values $B_h(P)$ and $B_h(r)$. ADT predicts binding on the log nM scale and we use 6.5 log nM for our binding threshold γ. Since our simulated data only uses 2-digit alleles we use the median ADT predicted binding across all observed 4-digit HLA alleles in Step that have the same first two digits to define $B_h(P)$ and $B_h(r)$ at the 2-digit level.

Under the null data simulation the test described above has the correct size for a nominal level of $\alpha = 0.05$ as shown in Table 5.3. As expected, power to detect simulated alternatives increases as the level of mutations in the vaccine group increases, also shown in Table 5.3. Notice, that when comparing tables 5.2 and 5.3 the average number of simulated mutations does not directly translate to the difference in mean number of potential escapes by treatment group. This is because a single mutation will often be counted as an escape by more than one HLA allele or the escape is counted more than once by overlapping reference peptides.

5.5 Analysis of Step trial

The method described in Section 5.3.2 was applied to the Gag protein using data from the Step trial. We use the same 64 infected trial participants as those defined in the analysis of the Gag protein in Rolland et al. study [71]. Specifically we use equation 5.4 to define

a potential CTL escape count for each infected participant. We run the analysis using the ADT binding prediction algorithm over a range of γ values representing binding thresholds of known CTL epitopes; the four γ values we considered are 5, 5.5, 6, and 6.5.

In the analysis by Roland et al [71] the authors detected a significant global difference in the peptide distances in Gag between treatment groups and a site specific difference at HXB2 position Gag 84. We are interested in whether or not our CTL specific method detects a global effect by using all 321 reference peptides $\mathcal{R}$ that have an unambiguous alignment across all 64 breakthrough sequences and the vaccine reference strain. We consider two additional sets of reference peptides to assess the overall contribution of Gag 84 to the observed effect; the first,

$$\mathcal{R}_{\setminus \mathrm{SL9}} \equiv \mathcal{R} \setminus \{\mathrm{SLYNTVATL}\}$$

eliminates the SLYNTVATL epitope from the reference peptides and the second,

$$\mathcal{R}_{\setminus 84} \equiv \mathcal{R} \setminus \{\mathrm{RSLYNTVAT}, \ldots, \mathrm{TLYCVHQKI}\}$$

eliminates all 9 reference peptides overlapping Gag 84.

5.5.1 Analysis of Step trial results

We see a global effect of treatment on the number of potential CTL escapes across the range of γ values considered (see Table 5.4) with an average of 11.32 potential escapes in the placebo group compared to 16.28 in the vaccine group for $\gamma = 6.5$. These averages are remarkably similar to the simulation study averages for the "High" level of mutation. When comparing the $\mathcal{R}_{\setminus \mathrm{SL9}}$ and $\mathcal{R}_{\setminus 84}$ results we see some attenuation of the mean escapes towards zero in both groups as well as P-values that are somewhat less significant but overall these results suggest that the overall effect is not driven solely by escape at position Gag 84.

The complementary analysis is to consider the reference peptide sets defined by SL9 alone $\mathcal{R}_{\mathrm{SL9}}$ and the reference peptide set consisting of all peptides that overlap Gag 84 $\mathcal{R}_{84}$. For these reference sets, depending on the level of γ, some subjects may not have a predicted binder among their set of HLA alleles to any of the reference peptides. In this case Y is by definition zero which has a different interpretation than the case where none

of predicted binders for a subject exceed the binding threshold necessary to be counted as a potential escape. In this case two valid analyses exist. We choose to analyze only those subjects with at least one predicted binder among the reference peptides. Table 5.5 shows the results for the reference peptide sets $\mathcal{R}_{\text{SL9}}$ and $\mathcal{R}_{84}$. For each set the table indicates the number of subjects included in the analysis.

These analyses show significant differences in the number of potential CTL escapes between treatment groups against the SL9 epitope for all values of γ analyzed. The analysis of $\mathcal{R}_{84}$ is more difficult to interpret because the value of γ changes the results. For $\gamma = 5$ or 5.5 there are a greater number of potential CTL escapes detected in the vaccine group compared to the placebo group and this difference is statistically significant at a level $\alpha = 0.05$ with P-values of 0.02 and 0.05 respectively. For $\gamma = 6$ or 6.5 there are still a greater number of potential CTL escapes detected in the vaccine group but P-values are greater than 0.2.

A table of the number of potential CTL escapes by treatment assignment is helpful for interpretation of the results of the $\mathcal{R}_{\text{SL9}}$ analysis.

	Placebo	Vaccine
0	12	11
1	5	10
2	0	6

The vaccine group has both a greater frequency of subjects with a potential CTL escape but also has several subjects with 2 alleles that contribute to their counts.

This method allows both the restriction of the reference peptide sets as seen in the various reference sets including or excluding the Gag 84 position. We can also restrict the alleles considered for a given analysis. For example, we can restrict the analysis to the HLA-A alleles and similarly the HLA-B alleles. Although there is not a clear biological motivation for these sub-analyses the results are show in Table 5.6.

Further sub-analyses focus on analyzing one allele or one reference peptide at a time.

For these analyses we report the results for a single γ value set at 6.5, the most liberal binding threshold we have used in the previous analyses. The diversity of HLA class I alleles makes a sub-analysis focusing on single 4-digit alleles meaningful only for a handful of alleles that are common enough to be represented by several subjects in both treatment arms. For the HLA specific analysis we consider the 6 alleles that are represented by at least 10 infected Step trial participants considered in the Rolland study. Due to small sample sizes and the discreteness of the data a permutation test, based on 1,000 permutations, was used to compare means between treatment groups rather than a t test (see Table 5.7). In 5 of the 6 alleles reported, the mean number of potential CTL escapes is higher in the vaccine group than in the placebo group. Also, the results give a sense of varying contribution to the overall averages seen in 5.4 across different alleles. Some alleles have more predicted binders to vaccine strain reference peptides than others reflected in the variation in means across alleles which is generally greater than the variation in means across treatment groups. For each of the 321 vaccine strain reference peptides in $\mathcal{R}$ we ran the analysis, again using a permutation test, with 1,000 permutations, to test for a difference in means between the treatment groups. In this analysis, only 4 of the 321 peptides have a P-value less than 0.1 (see Table 5.8). These HLA and peptide specific analyses highlight the utility of aggregating across sites and HLA alleles to detect a vaccine effect in this type of analysis.

5.6 Permutation analysis of PTE method

In this method the outcome depends on HLA and sequence. It seems intuitive that permutations of HLA and/or sequence would provide additional insight into these analyses. We demonstrate the use of permutation testing in this setting starting with a simple theoretical model then applying the methodology developed here to the Step data.

First consider a two sample problem with only a single two level covariate. Our two samples are defined by treatment assignment $Z \in \{0, 1\}$. The covariate, H, is a dichotomized HLA variable where $H = 0$ indicates weak HLA types that on average put less pressure on a virus than strong HLA types indicated by $H = 1$. Suppose we define a random variable

D which controls the amount of CTL escape as

$$\begin{aligned} D|H=0 &\sim N(0,1) \\ D|H=1 &\sim N(1,1). \end{aligned} \tag{5.6}$$

Define PTE in our hypothetical scenario as $Y = 11 + 6D$ which is intended to give escape counts similar to those seen in the Step Gag analysis although here we simulate a continuous PTE outcome rather than generate true count data.

Consider testing the null hypothesis that the dichotomized HLA covariate H is identically distributed between treatments,

$$H_0 : Pr(H = 1|Z = 1) = Pr(H = 1|Z = 0) \tag{5.7}$$

versus the alternative hypothesis that strong HLA types in the active treatment arm, $Z = 1$, and weak HLA types in the control arm, $Z = 0$, are both equally likely with probability $4/5$,

$$H_a : Pr(H = 1|Z = 1) = Pr(H = 0|Z = 0) = 4/5. \tag{5.8}$$

For this alternative, $E(Y|Z = 1) = 15.8$ and $E(Y|Z = 0) = 12.2$ which are similar to the observed mean PTE counts by treatment group seen in the analysis of the Step Gag data (see Section 5.5).

Since strong HLA types will generate more escape, we can test the null hypothesis using a t test. Alternatively, we consider a permutation approach using potential outcomes. In practice we use a binding prediction algorithm to define our outcome Y as described in the preceding sections. One thing that makes the CTL escape method unique is we have a type of potential outcome. Since binding prediction is defined for any combination of HLA and sequence we can create potential outcomes as in Rubin [73] where we "assign" HLA type. In our simple example we consider two potential outcomes; $D(1)$ is the outcome given the subject was assigned a strong HLA and $D(0)$ is the outcome give an assignment to a weak HLA. Continuing our example above, $D(1)$ and $D(0)$ define potential PTE outcomes $Y(1) = 11 + 6D(1)$ and $Y(0) = 11 + 6D(0)$.

5.6.1 Permutation tests

In our alternative testing framework based on permutations, we consider permuting HLA assignments rather than the typical approach of permuting treatment assignments. We follow the discussion of permutation testing by Romano [72]. Our data X take values in the sample space $\mathcal{X}$. Let $\mathbf{G}$ be a finite group of transformations such that for $g \in \mathbf{G}$, $gX \equiv g(X) \in \mathcal{X}$ and under the null hypothesis $gX \stackrel{d}{=} X$; under these conditions we refer to the distribution of X as invariant under the transformations in $\mathbf{G}$. In other treatments of permutation testing the invariance condition is referred to as exchangeability. Note that invariance states that the joint distribution of the probability law generating all of the observations is the same after the transformation g. Here we focus on transformations which are permutations of the covariates (e.g., HLA type).

A permutation test based on the set of transformations $\mathbf{G}$ proceeds as follows. Let $T(X)$ be any test statistic for testing H_0 and suppose $\mathbf{G}$ has M elements. For x in $\mathcal{X}$, let

$$T^{(1)}(x) \leq T^{(2)}(x) \leq \cdots \leq T^{(M)}(x) \tag{5.9}$$

be the ordered values of the set of statistics $T(gx)$ where $g \in \mathbf{G}$.

For any level $\alpha \in (0,1)$, we can define an exact two sided level α test as follows. Let

$$\begin{aligned} l &= \lceil \tfrac{M\alpha}{2} \rceil \\ r &= M - \lfloor \tfrac{M\alpha}{2} \rfloor \end{aligned} \tag{5.10}$$

where $\lceil M\alpha/2 \rceil$ denotes the smallest integer greater than or equal to $M\alpha/2$ and $\lfloor M\alpha/2 \rfloor$ denotes the largest integer less than or equal to $M\alpha/2$. Define $M^l(x)$ to be the number of indices $j \in 1, \ldots, M$ such that $T^{(j)}(x) < T^{(l)}(x)$. Similarly define $M^r(x)$ to be the number of indices j such that $T^{(j)}(x) > T^{(r)}(x)$ and define $M^0(x)$ to be the number of indices j such that either $T^{(j)}(x) = T^{(l)}(x)$ or $T^{(j)}(x) = T^{(r)}(x)$.

Let

$$a(x) = \frac{M\alpha - M^l(x) - M^r(x)}{M^0(x)} \tag{5.11}$$

then define the permutation test function as

$$\phi(x) = \begin{cases} 1 & T(x) < T^{(l)}(x) \text{ or } T(x) > T^{(r)}(x) \\ a(x) & T(x) = T^{(l)}(x) \text{ or } T(x) = T^{(r)}(x) \\ 0 & T^{(l)}(x) < T(x) < T^{(r)}(x). \end{cases} \tag{5.12}$$

Since $M^l(x) + M^r(x)$ and $M^0(x)$ count the number of times $\phi(gx)$ falls in the first and second cases of equation 5.12 respectively, we have

$$\sum_{g \in \mathbf{G}} \phi(gx) = M^l(x) + M^r(x) + a(x)M^0(x) = M\alpha. \tag{5.13}$$

Given $X \sim P$, we also have $gX \sim P$ since under the null hypothesis we assume that X is invariant to transformations in $\mathbf{G}$. Therefore,

$$\begin{aligned} M\alpha &= E_P\left[\sum_{g \in \mathbf{G}} \phi(gX)\right] \\ &= \sum_{g \in \mathbf{G}} E_P\left[\phi(gX)\right] \\ &= \sum_{g \in \mathbf{G}} E_P\left[\phi(X)\right] \\ &= M E_P\left[\phi(X)\right] \end{aligned} \tag{5.14}$$

demonstrating that ϕ is an exact test of level α.

In practice we must handle ties between $T(x)$ and T^l or T^r as well as the computational time needed to compute $T(gX)$ for each $g \in \mathbf{G}$ when M is large. To handle ties we typically define $a(x) = 0$ rather than using equation 5.11 which makes the test conservative in the presence of ties. For large M, we can use Monte Carlo draws from the set $\mathbf{G}$ rather than compute all M values $T^{(j)}$ in Equation 5.9.

5.6.2 Example using a two level HLA covariate

We continue by fully describing the sample space $\mathcal{X}$ in the example described in the introduction to Section 5.6. We will draw a total of 64 subjects with 39 active treatment $Z = 1$ and 25 control assignments $Z = 0$ mimicking the number of breakthrough infections used in the Step Gag analysis. Under the null hypothesis H_0 we assume that the strong HLA

types have a frequency of 0.5 in both treatment arms,

$$H_0 : Pr(H = 1|Z = 1) = Pr(H = 1|Z = 0) = 0.5. \tag{5.15}$$

The potential outcomes are distributed as follows

$$\begin{aligned} D(0) &\sim N(0,1) \\ D(1) &\sim N(1,1). \end{aligned} \tag{5.16}$$

Therefore, for each $i \in 1, \ldots, 64$, we have $Z_i, H_i, D_i(0)$, and $D_i(1)$. Our data X is the collection of 64 such observations. If we let $\mathbf{G}$ be the set of permutations of the HLA covariate, the probability law generating X described above is the same as the probability law generating gX where gX is simply X with the values of H permuted. As described in section 5.6.1, we can define an exact level α test based on any statistic $T(X)$.

Next we define four linear PTE models to compare our standard testing approach defined in Section 5.3.3 to the permutation testing approach.

Linear PTE Model 1: HLA mediated treatment effect

Define Y_i for $i \in 1, \ldots, 64$ as

$$Y_i = 11 + 6\left(\mathbf{I}_{[H_i=1]}D(1) + \mathbf{I}_{[H_i=0]}D(0)\right). \tag{5.17}$$

Without loss of generality we can assume that $Z_i = 0$ for $i \in 1, \ldots 25$ and $Z_i = 1$ for $i \in 26, \ldots, 64$. We define our test statistic $T(X)$ as the difference in means between treatment groups thus

$$T(X) = \frac{1}{39}\sum_{i=26}^{64} Y_i - \frac{1}{25}\sum_{i=1}^{25} Y_i = \overline{Y}_1 - \overline{Y}_0. \tag{5.18}$$

Detecting a treatment effect. We can directly compare the power to detect a treatment effect using the permutation test described above with an equal variance two-sample t test. Consider the alternative hypothesis

$$H_a : Pr(H = 1|Z = 1) = Pr(H = 0|Z = 1) = 4/5. \tag{5.19}$$

Note that the null hypothesis for the permutation test and the t test are different but both tests are testing for a treatment effect; we will use H_0^P and H_0^T respectively to distinguish the two null hypotheses. The null hypothesis for the permutation test

$$H_0^P : Pr(H = 1|Z = 1) = Pr(H = 1|Z = 0) \tag{5.20}$$

as defined above and the null hypothesis for the t test is

$$H_0^T : E(Y|Z = 1) = E(Y|Z = 0). \tag{5.21}$$

The power of the t test for rejecting H_0^T against the alternative H_a at a level of α is 0.57 while the power of the permutation test for rejecting H_0^P against the alternative H_a is 0.798. At first this may seem to contradict the uniformly most powerful (UMP) property of the t test however it is not too surprising in light of the fact that the tests are based on differing null hypotheses and the permutation test uses the potential outcome data which is ignored by the t test. A table of the power of each test for rejecting the respective null hypothesis for a variety of alternative hypotheses is shown in Table 5.9.

Linear PTE Model 2: Mixing HLA specific and non-HLA specific treatment effects

It is also interesting to compare the power of the two tests under a scenario that mixes a treatment specific HLA effect on CTL escape with a non-HLA associated treatment effect on CTL escape. Consider the following model for CTL escapes. Let

$$Y = 11 + m\mathbf{I}_{[Z=1]} + (6 - m)\mathbf{I}_{[H=1]} + \epsilon_1\mathbf{I}_{[H=1]} + \epsilon_0\mathbf{I}_{[H=0]}, \tag{5.22}$$

where ϵ_1 and ϵ_0 are independent mean zero normal distributions with standard deviation 6 and m is a mixing constant which defines the amounts of treatment specific HLA effect and non-HLA specific effect. When $m = 0$ the CTL effect that is completely mediated by HLA while when $m = 6$ the CTL effect is completely determined by treatment and is non-HLA specific. As before assume that $Pr(H = 1|Z = 1) = Pr(H = 0|Z = 0) = 4/5$ so we have differential HLA covariates by treatment assignment. The potential outcomes $Y(0)$ and $Y(1)$ are determined by HLA type as well as the error terms ϵ_0 and ϵ_1 in equation 5.22. Table 5.10 shows the power of each test for this model with $m \in 0, \ldots, 6$. As the mixing

coefficient increases, creating more non-HLA specific effect and less HLA specific effect, the t test gains power for rejecting H_0^T while the permutation test looses power for rejecting H_0^P. When $m = 6$ the permutation null hypothesis H_0^P is true and we see that the size of the permutation test is correct within Monte Carlo error for level $\alpha = 0.05$ in agreement with the theoretical permutation test results.

Linear PTE Model 3: HLA specific and sequence specific treatment effects

Next we consider a linear model of CTL escape which incorporates both HLA and sequence effects. In the this model we consider both the two level covariate H for HLA already considered along with a two level covariate for sequence S. Let

$$Y = 11 + m\mathbf{I}_{[H=1]} + (6-m)\mathbf{I}_{[S=1]} + \epsilon \tag{5.23}$$

where m is a mixing coefficient and the error term

$$\epsilon = \epsilon_{1,1}\mathbf{I}_{[H=1,S=1]} + \epsilon_{1,0}\mathbf{I}_{[H=1,S=0]} + \epsilon_{0,1}\mathbf{I}_{[H=0,S=1]} + \epsilon_{0,0}\mathbf{I}_{[H=0,S=0]} \tag{5.24}$$

is defined by $\epsilon_{1,1}, \epsilon_{1,0}, \epsilon_{0,1},$ and $\epsilon_{0,0}$ which are four independent mean zero normal distributions with standard deviation 6. The four combinations of the HLA and sequence covariates define four potential outcomes $Y(1,1), Y(1,0), Y(0,1),$ and $Y(0,0)$.

Under this model our null hypothesis for the permutation test is

$$H_0^P : Pr(H = 1|Z = 1) = Pr(H = 1|Z = 0) \text{ and } Pr(S = 1|Z = 1) = Pr(S = 1|Z = 0). \tag{5.25}$$

In this model we can consider permuting either HLA alone, sequence alone or both HLA and sequence for our permutation test. All three permutations are invariant transformations under H_0^P. We test the power of these three permutation tests against the alternative hypothesis

$$H_a : Pr(H = 1|Z = 1) = Pr(H = 0|Z = 0) = 4/5, Pr(S = 1|Z = 1) = Pr(S = 0|Z = 0) = 4/5 \tag{5.26}$$

for each mixing coefficient $m \in 0, \ldots, 6$.

Results are shown in figure 5.3. The t test is more powerful under all scenarios studied although the permutation test permuting both HLA and sequence is nearly as powerful. Power of the permutation tests that rely on permuting only HLA or sequence are driven by the amount of mixing.

Linear PTE Model 4: HLA drives sequence effect

A final example in our linear models study of permutation testing for CTL effects again employs two level covariates HLA and sequence. This model however considers the case where HLA drives sequence differences. We consider the same model as before considering only the case $m = 3$. What makes this example different is the alternative hypothesis,

$$\begin{aligned} H_a: \quad & Pr(H = 1|Z = 1) = Pr(H = 0|Z = 0) = p_1 \\ & \text{and} \\ & Pr(S = 1|H = 1) = Pr(S = 0|H = 0) = p_2. \end{aligned} \tag{5.27}$$

In the previous example p_1 was the fixed value $4/5$. Now we assess power over values of $p_1 \in [0.5, 1]$ and introduce the matching probability p_2 which we also vary over $[0.5, 1]$. In this example $H = 1$ or $S = 1$ tend to increase the CTL escape count while the probability that $S = 1$ is dependent on the probability that $H = 1$ through the probability p_2.

Results are shown in figure 5.4. In this model the t test is most powerful for detecting a treatment effect across most match probabilities with the permutation strategy which permutes both HLA and sequence nearly as powerful. As the match probability increases the power of these two test increase as well since the overall CTL effect increases in the model. When the match probability is $p_2 = 0.5$ the most powerful test is a the permutation test which permutes only the HLA covariate. In this strategy the power is constant across match probabilities since the added information due to sequence differences is not detected. Permuting sequences results in low power to detect a treatment effect until the match probability $p_2 = 1$ in which case this method is as powerful as the HLA permutation method since they are effectively equivalent methods in this case. When $p_1 = 0.5$ the null hypothesis is true and we see that for $\alpha = 0.05$ all four tests have the correct size.

Permutation Analysis of the Step Gag data

We analyze the Step Gag breakthrough sequences using the permutation analysis discussed in this section. In addition to three permutation strategies discussed in the linear PTE models (i.e., permuting HLA genotype and sequences either individually or separately) we consider four additional permutation strategies. For the sequence data we consider a strategy that permutes every aligned peptide separately wheras when permuting sequences we permute each row as a unit. In the Step Gag analysis we considered 321 reference peptides for 64 subjects. This creates a data matrix with dimensions 64×321; peptide permutation permutes each column separately. We consider a similar method with HLA alleles. In the linear PTE models we permuted HLA by permuting each genotype as a unit. A complementary strategy is to permute each allele separately. Finally, we consider permuting HLA and sequence within treatment assignment which can also be done at the allele and peptide level. For a permutation test to be valid we must consider whether or not we can define a null hypothesis in which the invariance condition is satisfied.

The null hypothesis and an invariance requirement for the three permutation strategies used in the linear PTE model examples apply here. One observation that should be made is that we may not be able to claim invariance under these strategies for the Step data because there may be HLA/sequence combinations that are more or less likely than others under the true data generating probability law. In fact, some combinations may be an impossibility. For permutation on the allele and/or peptide level this complication becomes exaggerated. For example, permuting overlapping sequences will likely create a set of peptides that could not come from the same sequence. Similarly, permuting alleles may produce genotypes that are rare in a population due to linkage disequilibrium. With these caveats we run the permutation analyses under these seven permutation strategies. In addition, we consider the standard permutation test which permutes treatment assignment. Table 5.11 list the P-values for each permutation strategy which range from 0.004 for the peptide permutation strategy to 0.06 for HLA permutation.

5.7 Discussion

In this chapter we have described a novel method for analyzing CTL effects based on breakthrough sequences from an HIV-1 vaccine trial. This method uses the potential outcome for an "assigned" HLA type derived from an MHC binding prediction algorithm. The potential outcomes allow us to better capture potential T-cell escape using the observed sequence variation for each reference K-mer. Incorporating HLA information into sieve analysis provides a unique challenge due to the high dimensional nature of the covariate yet using binding prediction and a form of potential outcomes we are able to exploit the richness of this covariate space to generate a method which is able to detect small effects using aggregated PTE counts accross many potential epitopes. We have also demonstrated how the method can be used locally to assess whether or not an observed effect is driven by a single local effect as we did in Section 5.5 using SLYNTVATL as an example.

The later half of this chapter focusses on assessing whether or not the effect can be linked to HLA genotype or sequence attributes that are differential across treatment. We began with an exploration of simple linear models which demonstrate that tests based on permutation of HLA and/or sequences are able to detect effects specific to one or both of these factors. This testing framework was applied to the Step trial data with the caveat that the invariance condition necessary to control the size of the test may not hold. Hutson and Wilding [34] propose a method for maintaining the exchangeability assumption in a non-randomized setting which may apply to our setting as well. In future work we would also like to expand the idea of potential outcomes in this setting to determine if we can interpret the results as causal effects in the framework of Rubin [73].

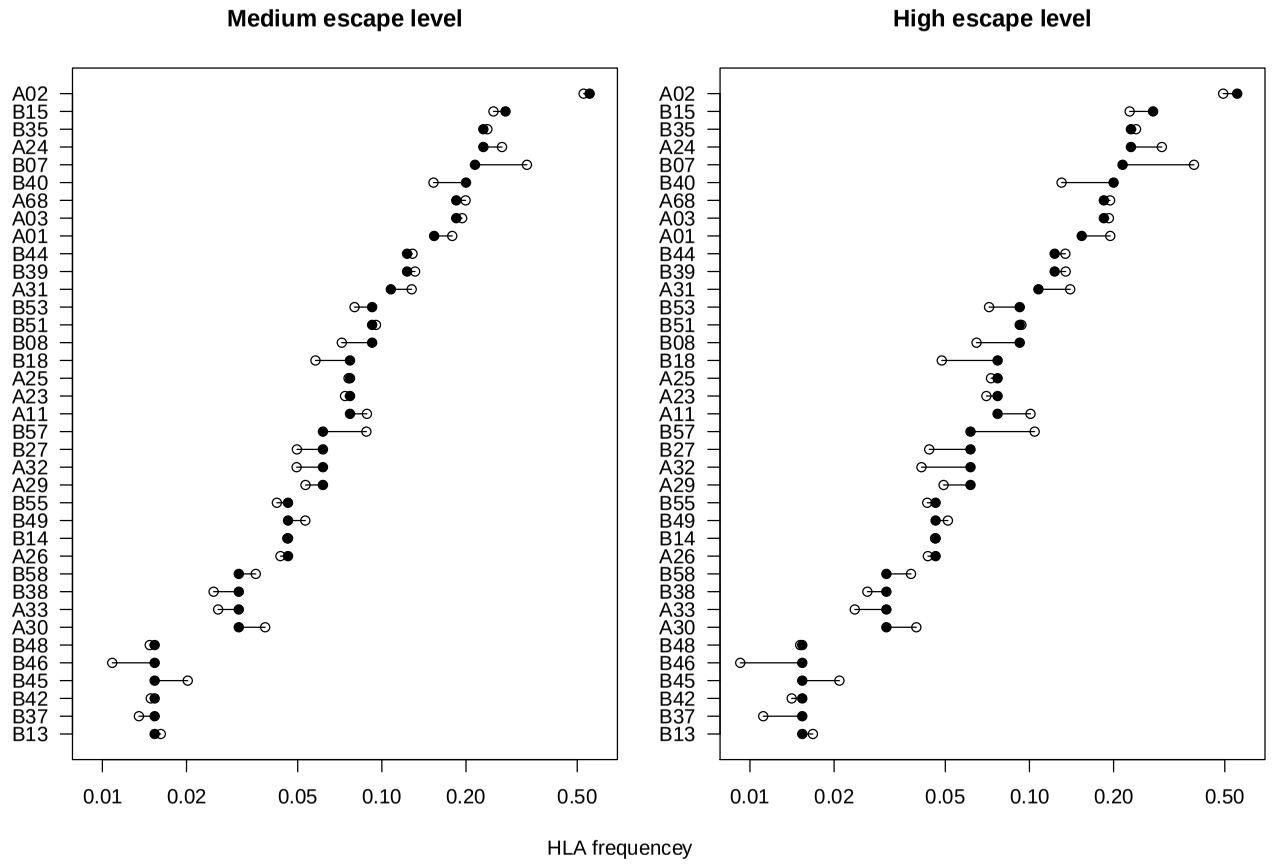

Figure 5.1: Simulation Study HLA frequency. In the simulation study placebo HLA alleles are sampled at the observed frequencies, shown as black dots, of the 64 participants study in Rolland et al [71] as are the HLA alleles for the vaccine group in the "Low" escape level simulations. For "Medium" and "High" escape level simulations the distribution of HLA alleles changes by design of the simulation study. The open dots in the left panel show the observed frequencies for each HLA allele in the vaccine group for the "Medium" escape level. Similarly the open dots in the right panel show the observed frequencies for each HLA allele in the vaccine group for the "High" escape level.

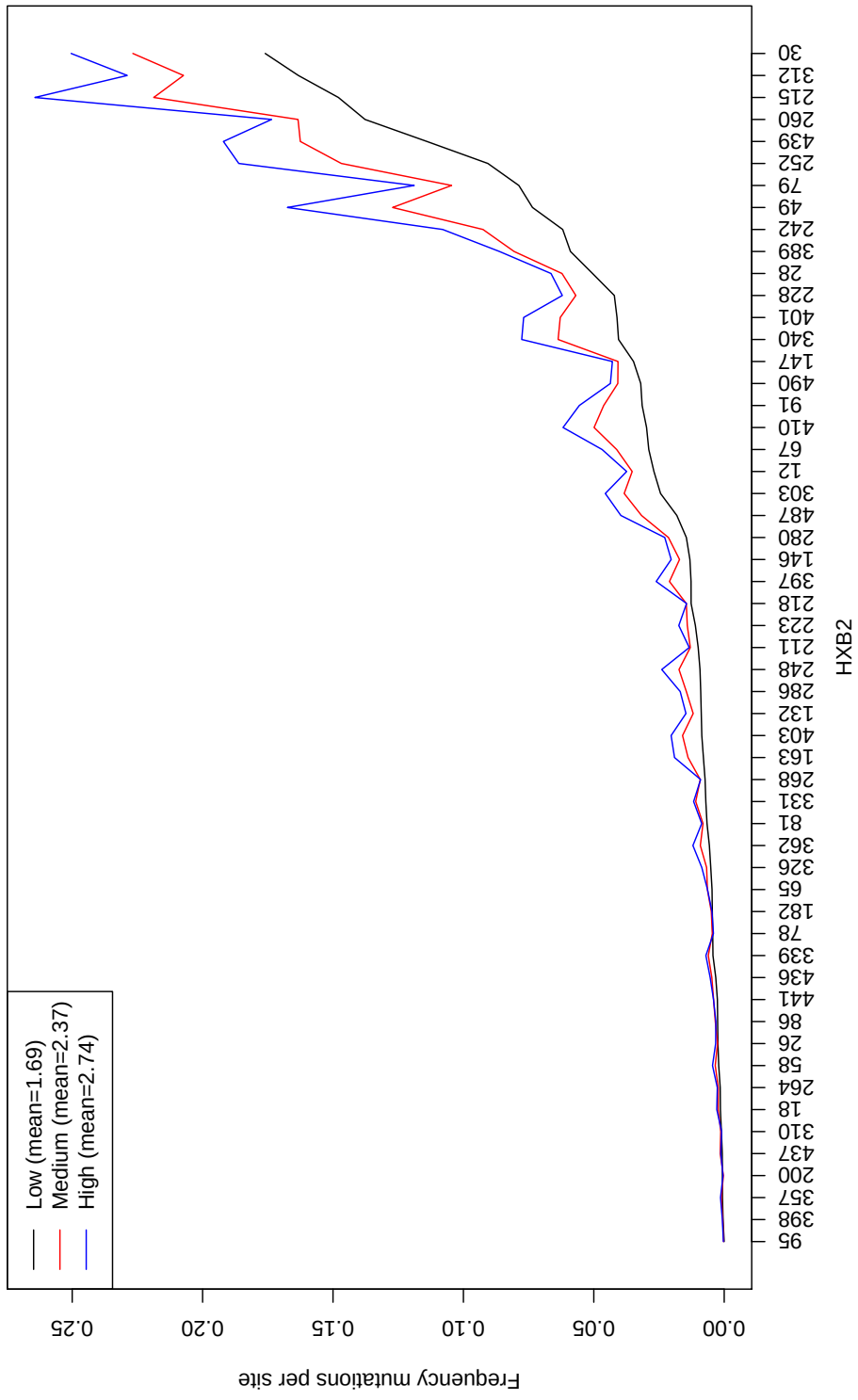

Figure 5.2: Simulation study sequence mutation frequency. All positions that were sites of simulated mutations within the vaccine sequences are listed on the x-axis in the order of the frequency of mutation. The three lines show the frequency of mutation at each position for each of the levels of simulated mutation; "Low", "Medium", and "High". The means given in the legend state the overall mutation rate per sequence for each simulated mutation level.

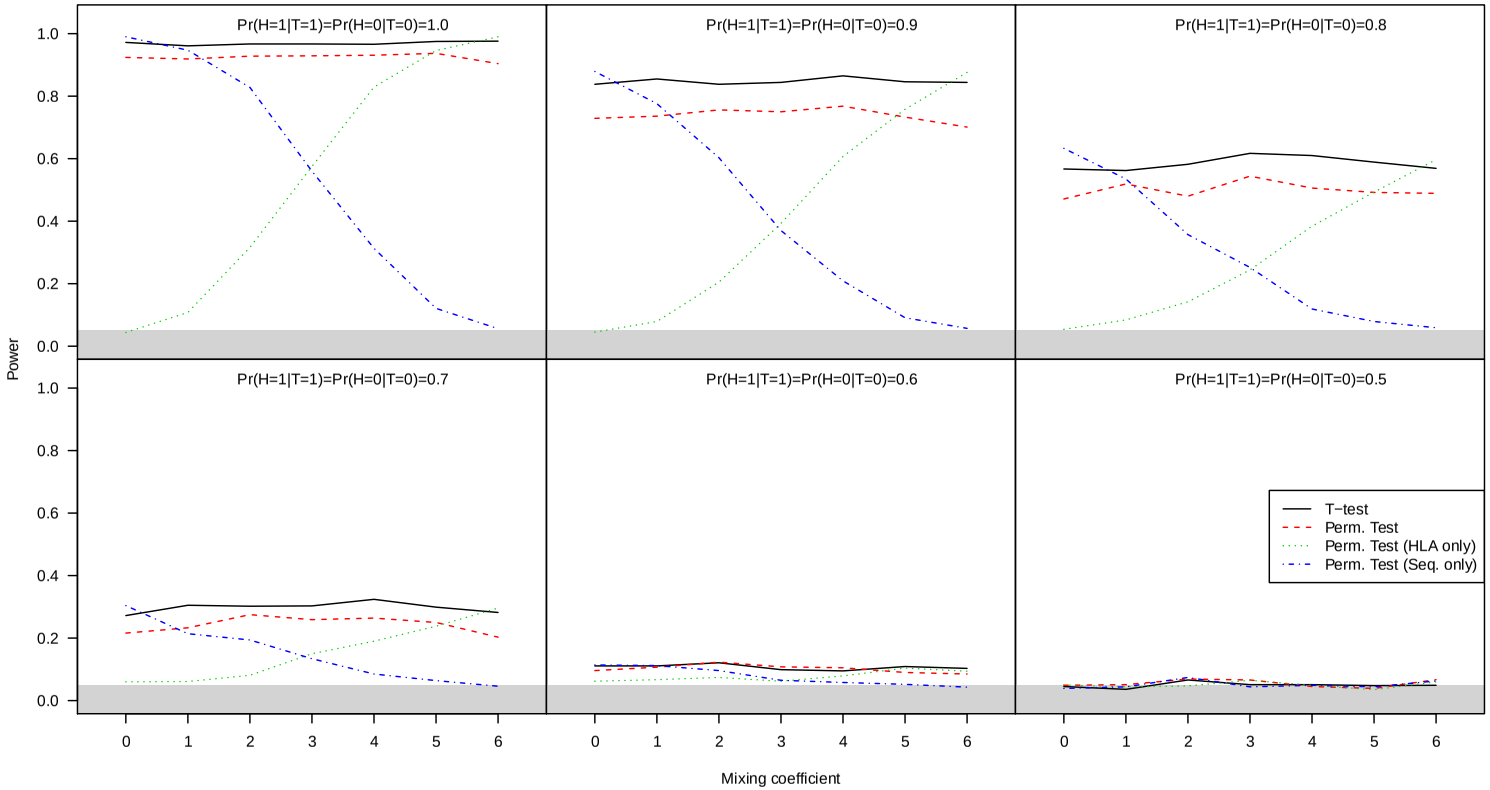

Figure 5.3: Power analysis for Linear PTE Model 3 (see Section 5.6.2). Probabilities from 1 to 0.5 defining both the HLA and sequence distributions by treatment assignment are used to generate each panel. The y-axis shows the power for level $\alpha = 0.05$ as the mixing coefficient m ranges from 0 to 6. When the mixing coefficient is $m = 0$ the CTL effect is completely determined by the sequence covariate while for $m = 6$ the CTL effect is completely determined by the HLA covariate. Each of the 4 lines shows the power for the t test and three variants of permutation test as given in the legend. In lower right panel, with probability 0.5, the null hypothesis is satisfied and the plot shows the size of each test is correct for $\alpha = 0.05$. The shaded rectangle at the bottom of each panel indicates power of 0.05 or less.

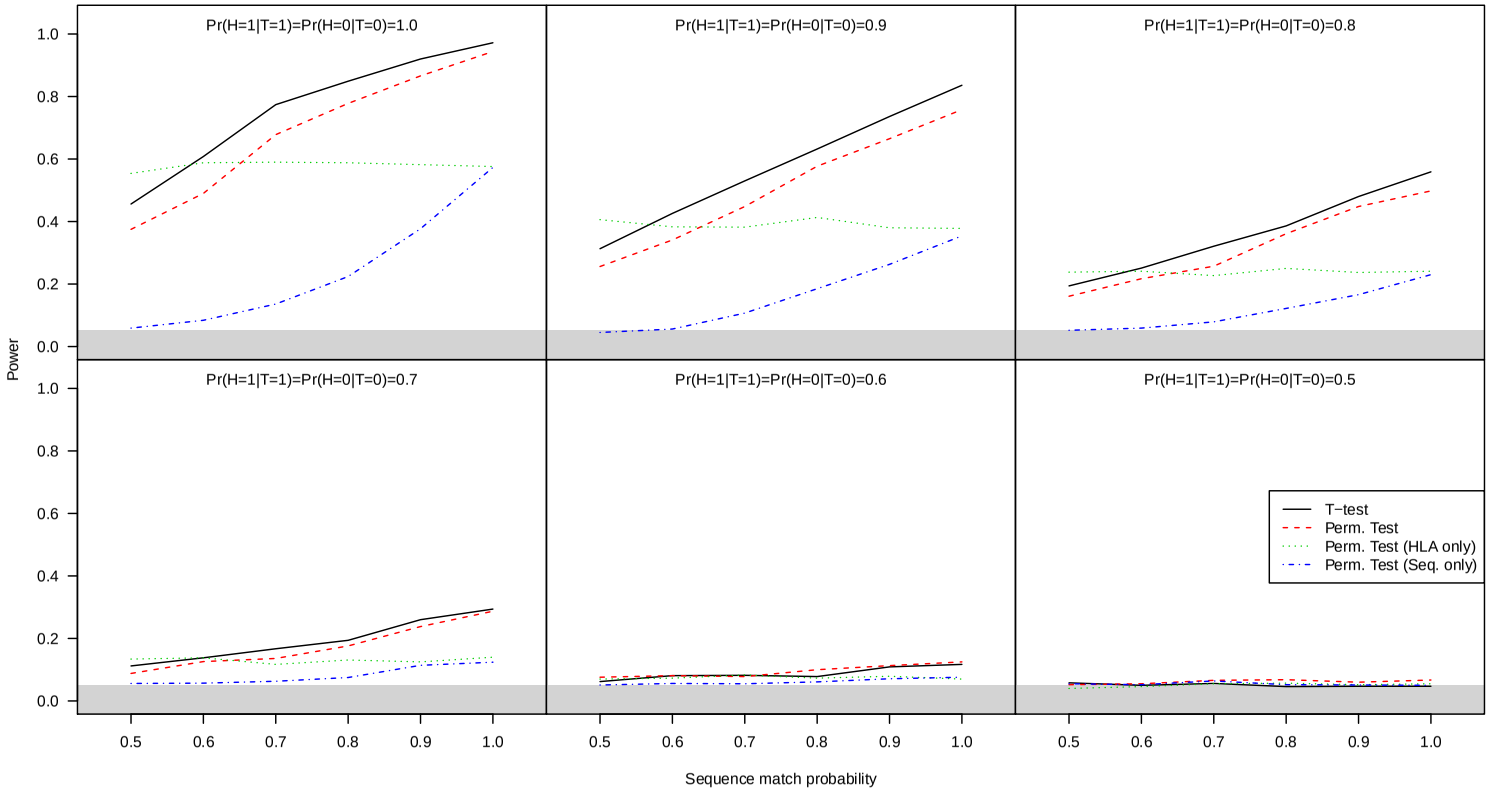

Figure 5.4: Power analysis for Linear PTE Model 4 (see Section 5.6.2). Probabilities from 1 to 0.5 defining both the HLA distributions by treatment assignment are used to generate each panel. The y-axis shows the power for level $\alpha = 0.05$ as the mixing sequence match probability p_2 ranges from 0.5 to 1. When the matching probability $p_2 = 0.5$, the sequence distribution is independent of HLA and treatment assignment. When the matching probability $p_2 = 1$, the sequence distribution has correlation 1 with the HLA covariate. Each of the 4 lines shows the power for the t test and three variants of permutation test as given in the legend. In lower right panel, with probability 0.5, the null hypothesis is satisfied and the plot shows the size of each test is correct for $\alpha = 0.05$. The shaded rectangle at the bottom of each panel indicates power of 0.05 or less.

	Allele count		
	2	3	4
Low	0.3	11.5	88.1
Medium	0.1	8.5	91.3
High	0.1	7.1	92.8

Table 5.1: Simulation study HLA count frequency. For each level of mutation in the simulation study the percentage of vaccine subjects with the given allele count are shown. By design of the study, allele frequency counts in the placebo group are the same as the vaccine group for the "Low" level of mutation. Each simulated trial participant has either 2, 3, or 4 unique HLA-A and -B alleles. For simulated HLA genotypes with 2 alleles the simulated HLA genotype is homozygotic at both the HLA-A and -B loci. For simulated genotypes with 3 alleles either the HLA-A or -B loci is homozygotic. If the count is 4, all alleles are distinct.

	Sequence		**Number of sites**	
	Mean	SD	Mean	SD
Low	1.69	1.29	24.71	2.85
Medium	2.37	1.27	28.35	2.60
High	2.74	1.23	29.80	2.68

Table 5.2: Simulation study sequence mutation statistics. The mean and standard deviation for the number of mutation per sequence are given in the "Sequence" column" for each level of mutation. Similarly the "Number of sites" column shows the mean and standard deviation for the number of mutated sites within the vaccine group across all 1,000 simulated data sets.

	Placebo		**Vaccine**			**95% CI**	
	Mean	SD	Mean	SD	Power	Lower	Upper
Null	11.75	1.19	11.73	0.91	5.2	3.94	6.81
Low	11.55	1.17	14.22	1.04	37.1	34.11	40.19
Medium	11.37	1.19	15.43	1.05	68.7	65.71	71.55
High	11.41	1.21	16.09	1.11	77.4	74.65	79.93

Table 5.3: Simulation study power analysis. The first column shows the mean value of the mean count of potential escapes in the placebo group across the 1,000 simulations for each of the levels of mutation plus the null distribution. The standard deviation of the means is also show. Similarly the mean and standard deviation for the mean counts in the vaccine group across the simulations are shown. For $\alpha = 0.05$, the percentage of simulated data sets with a p-value less than α are shown along with a 95% confidence interval.

	$\mathcal{R}$			$\mathcal{R}_{\setminus \mathbf{SL9}}$			$\mathcal{R}_{\setminus \mathbf{84}}$		
γ	Plac	Vacc	P-val.	Plac	Vacc	P-val.	Plac	Vacc	P-val.
5.0	2.88	4.56	0.04	2.76	4.15	0.07	2.60	3.85	0.08
5.5	5.08	7.59	0.05	4.96	7.13	0.08	4.44	6.54	0.08
6.0	7.68	11.49	0.03	7.52	10.92	0.04	6.64	10.08	0.03
6.5	11.32	16.28	0.04	11.12	15.72	0.05	10.16	14.72	0.04

Table 5.4: Step Gag analysis. We ran the CTL binding escape analysis for three sets of reference peptides as described in Section 5.5. Each analysis was run for $\gamma = 5, 5.5, 6,$ and 6.5. For each peptide reference set the observed mean number of potential CTL escapes is shown for each treatment group along with the P-value from a t test for a difference in group means.

	$\mathcal{R}_{\mathbf{SL9}}$				$\mathcal{R}_{\mathbf{84}}$			
γ	N	Plac	Vacc	P-val.	N	Plac	Vacc	P-val.
5.0	13,22	0.23	0.73	0.0100	18,31	0.39	0.90	0.0193
5.5	13,23	0.23	0.78	0.0054	24,36	0.67	1.14	0.0489
6.0	16,26	0.25	0.85	0.0032	25,39	1.04	1.41	0.2267
6.5	17,27	0.29	0.81	0.0088	25,39	1.16	1.56	0.2211

Table 5.5: Step analysis Gag 84. We ran the CTL binding escape analysis for the single reference peptide SL9 $\mathcal{R}_{\text{SL9}}$ as well as the set of 9 peptides overlapping Gag 84 $\mathcal{R}_{84}$. Each analysis was run for $\gamma = 5, 5.5, 6, \text{and } 6.5$. For each peptide reference set, the number of subjects in the placebo group and the vaccine group that have at least one HLA allele that binds to at least one reference peptide with an affinity at least γ are shown in the N column. The observed mean number of potential CTL escapes is shown for each treatment group along with the P-value from a t test for a difference in group means.

	HLA-A			**HLA-B**		
γ	Plac	Vacc	P-val.	Plac	Vacc	P-val.
5.0	1.56	2.54	0.0095	1.32	2.03	0.2977
5.5	2.36	4.15	0.0003	2.72	3.44	0.5213
6.0	3.76	6.00	0.0003	3.92	5.49	0.3033
6.5	5.32	8.05	0.0015	6.00	8.23	0.2927

Table 5.6: Analysis by HLA-A and HLA-B. We ran the CTL binding escape analysis separately for the HLA-A and HLA-B loci. Each analysis was run for $\gamma = 5, 5.5, 6, \text{and } 6.5$. For each loci, the observed mean number of potential CTL escapes is shown for each treatment group along with the P-value from a t test for a difference in group means.

Allele	N	Plac	Vacc	P-val.
HLA-A*02:01	9,18	4.22	5.72	0.12
HLA-A*24:02	5,8	5.00	7.62	0.06
HLA-B*07:02	4,9	0.50	1.11	0.49
HLA-A*68:01	6,6	3.50	2.67	0.32
HLA-A*03:01	5,7	1.40	2.00	0.48
HLA-B*15:01	7,4	0.29	1.00	0.09

Table 5.7: Analysis of most common HLA alleles. We ran the CTL binding escape analysis separately for the 6 most common HLA-A and HLA-B alleles. Each analysis was run for $\gamma = 6.5$. For each allele, the number of placebo and vaccine participants is shown under the heading N. The observed mean number of potential CTL escapes is shown for each treatment group along with the P-value from a permutation test, based on 1,000 permutations, for a difference in group means.

HXB2	Peptide	N	Plac	Vacc	P-val.
77	SLYNTVATL	17,27	0.29	0.81	0.02
234	SDIAGTTST	7,10	0.00	0.60	0.08
247	IGWMTNNPP	10,19	0.00	0.42	0.07
433	FLGKIWPSH	18,27	0.17	0.56	0.07

Table 5.8: Analysis of individual vaccine strain reference peptides. We ran the CTL binding escape analysis separately for each of the 321 vaccine strain reference peptides $\mathcal{R}$. Each analysis was run for $\gamma = 6.5$. For each peptide, the number of placebo and vaccine participants is shown under the heading N. The observed mean number of potential CTL escapes is shown for each treatment group along with the P-value from a permutation test, based on 1,000 permutations, for a difference in group means. The 4 peptides with a P-value less than 0.1 are reported in this table.

	Power	
$Pr(H = 1\|T = 1) = Pr(H = 0\|T = 0)$	T-test	Perm. Test
0.9	0.849	0.976
0.8	0.573	0.818
0.7	0.298	0.481
0.6	0.120	0.149
0.5	0.047	0.052

Table 5.9: Comparison of power between t test and permutation testing in linear PTE model 1 (see Section 5.6.2). Power for a level $\alpha = 0.05$ of each test for probabilities between ranging from 0.9 to 0.5 are shown where the probability defines the distribution of the HLA covariate in each treatment group. When the probability is 0.5 the HLA covariate has the same distribution in each group so the null hypothesis is satisfied and the power gives the size of the test.

	Power	
Mixing coefficient	T-test	Perm. Test
0	0.572	0.793
1	0.679	0.720
2	0.778	0.567
3	0.856	0.361
4	0.899	0.209
5	0.948	0.097
6	0.966	0.055

Table 5.10: Comparison of power between t test and permutation testing in linear PTE model 2 (see Section 5.6.2). Power for a level $\alpha = 0.05$ of each test for the mixing coefficient m ranging from 0 to 6. The mixing coefficient determines how much of the CTL effect is mediated by HLA specific and non-HLA specific CTL effects. When $m = 0$ all of the effect is non-HLA specific and power gives the size of the permutation test.

Permutation strategy	P-value
Treatment	0.024
HLA & Sequence	0.036
HLA	0.060
Sequence	0.024
HLA & Sequence within Treatment	0.034
Alleles	0.044
Peptides	0.004
Alleles & Peptides within Treatment	0.050

Table 5.11: The P-value for each permutation strategy is shown.

BIBLIOGRAPHY

[1] J. Arthos, C. Cicala, E. Martinelli, K. Macleod, D. Van Ryk, D. Wei, Z. Xiao, T. D. Veenstra, T. P. Conrad, R. A. Lempicki, S. McLaughlin, M. Pascuccio, R. Gopaul, J. McNally, C. C. Cruz, N. Censoplano, E. Chung, K. N. Reitano, S. Kottilil, D. J. Goode, and A. S. Fauci. HIV-1 envelope protein binds to and signals through integrin alpha4beta7, the gut mucosal homing receptor for peripheral t cells. *Nat. Immunol.*, 9(3):301309, March 2008.

[2] YA Ban, BE Correia, M Holmes, E Boni, N Sather, C Bretz, O Kalyuzhniy, C Xu, D Baker, L Stamatatos, R Strong, and W Schief. P05-09. 4e10 epitope-scaffolds mimic the antibody-bound epitope conformation and block neutralization by sera from rare HIV+ individuals. *Retrovirology*, 6(Suppl 3):P85, October 2009. PMID: null PMCID: PMC2767990.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, January 1995. ArticleType: research-article / Full publication date: 1995 / Copyright 1995 Royal Statistical Society.

[4] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, August 2001. ArticleType: research-article / Full publication date: Aug., 2001 / Copyright 2001 Institute of Mathematical Statistics.

[5] Mattia Bonsignori, Kwan-Ki Hwang, Xi Chen, Chun-Yen Tsao, Lynn Morris, Elin Gray, Dawn J. Marshall, John A. Crump, Saidi H. Kapiga, Noel E. Sam, Faruk Sinangil, Marie Pancera, Yang Yongping, Baoshan Zhang, Jiang Zhu, Peter D. Kwong, Sijy O'Dell, John R. Mascola, Lan Wu, Gary J. Nabel, Sanjay Phogat, Michael S. Seaman, John F. Whitesides, M. Anthony Moody, Garnett Kelsoe, Xinzhen Yang, Joseph Sodroski, George M. Shaw, David C. Montefiori, Thomas B. Kepler, Georgia D. Tomaras, S. Munir Alam, Hua-Xin Liao, and Barton F. Haynes. Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors. *Journal of Virology*, 85(19):9998–10009, October 2011.

[6] Patrick Breheny and Jian Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369–380, July 2009. PMID: 20640242.

[7] Zabrina L. Brumme, Mina John, Jonathan M. Carlson, Chanson J. Brumme, Dennison Chan, Mark A. Brockman, Luke C. Swenson, Iris Tao, Sharon Szeto, Pamela

Rosato, Jennifer Sela, Carl M. Kadie, Nicole Frahm, Christian Brander, David W. Haas, Sharon A. Riddler, Richard Haubrich, Bruce D. Walker, P. Richard Harrigan, David Heckerman, and Simon Mallal. HLA-Associated immune escape pathways in HIV-1 subtype b gag, pol and nef proteins. *PLoS ONE*, 4(8):e6687, August 2009.

[8] Susan P. Buchbinder, Devan V. Mehrotra, Ann Duerr, Daniel W. Fitzgerald, Robin Mogg, David Li, Peter B. Gilbert, Javier R. Lama, Michael Marmor, Carlos del Rio, M. Juliana McElrath, Danilo R. Casimiro, Keith M. Gottesdiener, Jeffrey A. Chodakewitz, Lawrence Corey, and Michael N. Robertson. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the step study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet*, 372(9653):1881–1893, November 2008. PMID: 19012954 PMCID: PMC2721012.

[9] Bing Chen, Erik M. Vogan, Haiyun Gong, John J. Skehel, Don C. Wiley, and Stephen C. Harrison. Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature*, 433(7028):834–841, February 2005.

[10] Karen N Conneely and Michael Boehnke. So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *American journal of human genetics*, 81(6):1158–1168, December 2007. PMID: 17966093.

[11] G. Dahmen and A. Ziegler. Generalized estimating equations in controlled clinical trials: Hypotheses testing. *Biometrical Journal*, 46(2):214232, 2004.

[12] P. Edlefsen, T. Hertz, C. Magaret, A deCamp, and Gilbert P. Sieve analysis of RV144. In *AIDS Vaccine*, Bangkok, Thailand, 2011.

[13] Bradley Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, March 2007.

[14] Jos Esparza. A brief history of the global effort to develop a preventive HIV vaccine. *Vaccine*, 31(35):3502–3518, August 2013. PMID: 23707164.

[15] Michael P. Fay and Barry I. Graubard. Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics*, 57(4):1198–1206, December 2001. ArticleType: research-article / Full publication date: Dec., 2001 / Copyright 2001 International Biometric Society.

[16] Neil M Flynn, Donald N Forthal, Clayton D Harro, Franklyn N Judson, Kenneth H Mayer, and Michael F Para. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *The Journal of infectious diseases*, 191(5):654–665, March 2005. PMID: 15688278.

[17] Dean Follmann, Michael Proschan, and Eric Leifer. Multiple outputation: inference for complex clustered data by averaging analyses from independent data. *Biometrics*, 59(2):420–429, June 2003. PMID: 12926727.

[18] Ron A M Fouchier and Derek J Smith. Use of antigenic cartography in vaccine seed strain selection. *Avian diseases*, 54(1 Suppl):220–223, March 2010. PMID: 20521635.

[19] Mark Gerstein, Erik L.L. Sonnhammer, and Cyrus Chothia. Volume changes in protein evolution. *Journal of Molecular Biology*, 236(4):1067–1078, March 1994.

[20] Peter Gilbert, Steve Self, Malla Rao, Abdollah Naficy, and John Clemens. Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. *Journal of Clinical Epidemiology*, 54(1):68–85, January 2001.

[21] Peter Gilbert, Maggie Wang, Terri Wrin, Chris Petropoulos, Marc Gurwith, Faruk Sinangil, Patricia D'Souza, Isaac R. Rodriguez-Chavez, Allan DeCamp, Mike Giganti, Phillip W. Berman, Steve G. Self, and David C. Montefiori. Magnitude and breadth of a nonprotective neutralizing antibody response in an efficacy trial of a candidate HIV-1 gp120 vaccine. *Journal of Infectious Diseases*, 202(4):595–605, August 2010.

[22] Peter B. Gilbert, Vladimir A. Novitsky, Monty A. Montano, and Max Essex. An efficient test for comparing sequence diversity between two populations. *Journal of Computational Biology*, 8(2):123–139, April 2001.

[23] Peter B. Gilbert, Steven G. Self, and Mark A. Ashby. Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics*, 54(3):799–814, 1998. ArticleType: research-article / Full publication date: Sep., 1998 / Copyright 1998 International Biometric Society.

[24] Peter B. Gilbert, Chunyuan Wu, and David V. Jobes. Genome scanning tests for comparing amino acid sequences between groups. *Biometrics*, 64(1):198207, 2008.

[25] S. Gnanakaran, Tanmoy Bhattacharya, Marcus Daniels, Brandon F. Keele, Peter T. Hraber, Alan S. Lapedes, Tongye Shen, Brian Gaschen, Mohan Krishnamoorthy, Hui Li, Julie M. Decker, Jesus F. Salazar-Gonzalez, Shuyi Wang, Chunlai Jiang, Feng Gao, Ronald Swanstrom, Jeffrey A. Anderson, Li-Hua Ping, Myron S. Cohen, Martin Markowitz, Paul A. Goepfert, Michael S. Saag, Joseph J. Eron, Charles B. Hicks, William A. Blattner, Georgia D. Tomaras, Mohammed Asmal, Norman L. Letvin, Peter B. Gilbert, Allan C. DeCamp, Craig A. Magaret, William R. Schief, Yih-En Andrew Ban, Ming Zhang, Kelly A. Soderberg, Joseph G. Sodroski, Barton F. Haynes, George M. Shaw, Beatrice H. Hahn, and Bette Korber. Recurrent signature patterns in HIV-1 b clade envelope glycoproteins associated with either early or chronic infections. *PLoS Pathog*, 7(9):e1002209, 2011.

[26] Paul G. Gottschalk and John R. Dunn. The five-parameter logistic: A characterization and comparison with the four-parameter logistic. *Analytical Biochemistry*, 343(1):54–65, August 2005.

[27] Javier Guenaga, Pia Dosenovic, Gilad Ofek, David Baker, William R. Schief, Peter D. Kwong, Gunilla B. Karlsson Hedestam, and Richard T. Wyatt. Heterologous epitope-scaffold PrimeBoosting immuno-focuses b cell responses to the HIV-1 gp41 2F5 neutralization determinant. *PLoS ONE*, 6(1):e16074, January 2011.

[28] Clayton D Harro, Franklyn N Judson, Geoffrey J Gorse, Kenneth H Mayer, Jay R Kostman, Stephen J Brown, Beryl Koblin, Michael Marmor, Bradford N Bartholow, and Vladimir Popovic. Recruitment and baseline epidemiologic profile of participants in the first phase 3 HIV vaccine efficacy trial. *Journal of acquired immune deficiency syndromes (1999)*, 37(3):1385–1392, November 2004. PMID: 15483468.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2nd ed. 2009. corr. 3rd printing 5th printing. edition, February 2009.

[30] B. F. Haynes, P. B. Gilbert, M. J. McElrath, S. Zolla-Pazner, G. D. Tomaras, S. M. Alam, D. T. Evans, D. C. Montefiori, C. Karnasuta, R. Sutthent, H. X. Liao, A. L. DeVico, G. K. Lewis, C. Williams, A. Pinter, Y. Fong, H. Janes, A. DeCamp, Y. Huang, M. Rao, E. Billings, N. Karasavvas, M. L. Robb, V. Ngauy, M. S. de Souza, R. Paris, G. Ferrari, R. T. Bailer, K. A. Soderberg, C. Andrews, P. W. Berman, N. Frahm, S. C. De Rosa, M. D. Alpert, N. L. Yates, X. Shen, R. A. Koup, P. Pitisuttithum, J. Kaewkungwal, S. Nitayaphan, S. Rerks-Ngarm, N. L. Michael, and J. H. Kim. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *N. Engl. J. Med.*, 366(14):12751286, April 2012.

[31] Elaine B. Hoffman, Pranab K. Sen, and Clarice R. Weinberg. Within-cluster resampling. *Biometrika*, 88(4):1121–1134, December 2001. ArticleType: research-article / Full publication date: Dec., 2001 / Copyright 2001 Biometrika Trust.

[32] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, January 1979. ArticleType: research-article / Full publication date: 1979 / Copyright 1979 Board of the Foundation of the Scandinavian Journal of Statistics.

[33] Yunda Huang, Peter B. Gilbert, David C. Montefiori, and Steve G. Self. Simultaneous evaluation of the magnitude and breadth of a left and right censored multivariate response, with application to HIV vaccine development. *Statistics in biopharmaceutical research*, 1(1):81–91, February 2009. PMID: 20072667 PMCID: PMC2805400.

[34] Alan D. Hutson and Gregory E. Wilding. Maintaining the exchangeability assumption for a two-group permutation test in the non-randomized setting. *Journal of Applied Statistics*, 39(7):1593–1603, 2012.

[35] Rob J. Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):pp. 361–365, 1996.

[36] H. Janes, N. Frahm, A. DeCamp, M. Rolland, E. Gabriel, J. Wolfson, T. Hertz, E. Kallas, P. Goepfert, D. P. Friedrich, L. Corey, J. I. Mullins, M. J. McElrath, and P. Gilbert. MRKAd5 HIV-1 gag/pol/nef vaccine-induced t-cell responses inadequately predict distance of breakthrough HIV-1 sequences to the vaccine or viral load PLoS ONE. *PLoS ONE*, 2012. Accepted.

[37] Xia Jin, Mark J Newman, Stephen De-Rosa, Cristine Cooper, Evan Thomas, Michael Keefer, Jonathan Fuchs, William Blattner, Brian D Livingston, Denise M McKinney, Elizabeth Noonan, Allan Decamp, Olivier D Defawe, and Margaret Wecker. A novel HIV t helper epitope-based vaccine elicits cytokine-secreting HIV-specific CD4+ t cells in a phase i clinical trial in HIV-uninfected adults. *Vaccine*, 27(50):7080–7086, November 2009. PMID: 19786145.

[38] Alla Katsnelson. Strategy to fight HIV shapes up. *Nature News*, September 2010.

[39] Gran Kauermann and Raymond J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, December 2001. ArticleType: research-article / Full publication date: Dec., 2001 / Copyright 2001 American Statistical Association.

[40] David R. Kaufman, Fusheng Li, Ashley N. Cruz, Steven G. Self, and Dan H. Barouch. Focus and breadth of cellular immune responses elicited by a heterologous insert prime-boost vaccine regimen in rhesus monkeys. *Vaccine*, 30(3):506–509, January 2012.

[41] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Wiley, New York, 1990.

[42] A. Lapedes and R. Farber. The geometry of shape space: Application to influenza. *JOURNAL OF THEORETICAL BIOLOGY*, 212:57–70, 2001. ID: 209446531.

[43] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, February 2004.

[44] Jan de Leeuw and Patrick Mair. Multidimensional scaling using majorization: SMACOF in r. *Journal of Statistical Software31*, 31(3), 2009.

[45] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, April 1986. ArticleType: research-article / Full publication date: Apr., 1986 / Copyright 1986 Biometrika Trust.

[46] Hua-Xin Liao, Mattia Bonsignori, S.Munir Alam, JasonS. McLellan, GeorgiaD. Tomaras, M.Anthony Moody, DanielM. Kozink, Kwan-Ki Hwang, Xi Chen, Chun-Yen Tsao, Pinghuang Liu, Xiaozhi Lu, RobertJ. Parks, DavidC. Montefiori, Guido Ferrari, Justin Pollara, Mangala Rao, KristinaK. Peachman, Sampa Santra, NormanL. Letvin, Nicos Karasavvas, Zhi-Yong Yang, Kaifan Dai, Marie Pancera, Jason Gorman, Kevin Wiehe, NathanI. Nicely, Supachai Rerks-Ngarm, Sorachai Nitayaphan, Jaranit Kaewkungwal, Punnee Pitisuttithum, James Tartaglia, Faruk Sinangil, JeromeH. Kim, NelsonL. Michael, ThomasB. Kepler, PeterD. Kwong, JohnR. Mascola, GaryJ. Nabel, Abraham Pinter, Susan Zolla-Pazner, and BartonF. Haynes. Vaccine induction of antibodies against a structurally heterogeneous site of immune pressure within HIV-1 envelope protein variable regions 1 and 2. *Immunity*, 38(1):176–186, January 2013.

[47] Bing Lu, John S Preisser, Bahjat F Qaqish, Chirayath Suchindran, Shrikant I Bangdiwala, and Mark Wolfson. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63(3):935–941, September 2007. PMID: 17825023.

[48] L A Mancl and T A DeRouen. A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57(1):126–134, March 2001. PMID: 11252587.

[49] J. R. Mascola, P. D'Souza, P. Gilbert, B. H. Hahn, N. L. Haigwood, L. Morris, C. J. Petropoulos, V. R. Polonis, M. Sarzotti, and D. C. Montefiori. Recommendations for the design and use of standard virus panels to assess neutralizing antibody responses elicited by candidate human immunodeficiency virus type 1 vaccines. *J. Virol.*, 79(16):1010310107, August 2005.

[50] J. S. McLellan, M. Pancera, C. Carrico, J. Gorman, J. P. Julien, R. Khayat, R. Louder, R. Pejchal, M. Sastry, K. Dai, S. O'Dell, N. Patel, S. Shahzad-ul Hussan, Y. Yang, B. Zhang, T. Zhou, J. Zhu, J. C. Boyington, G. Y. Chuang, D. Diwanji, I. Georgiev, Y. D. Kwon, D. Lee, M. K. Louder, S. Moquin, S. D. Schmidt, Z. Y. Yang, M. Bonsignori, J. A. Crump, S. H. Kapiga, N. E. Sam, B. F. Haynes, D. R. Burton, W. C. Koff, L. M. Walker, S. Phogat, R. Wyatt, J. Orwenyo, L. X. Wang, J. Arthos, C. A. Bewley, J. R. Mascola, G. J. Nabel, W. R. Schief, A. B. Ward, I. A. Wilson, and P. D. Kwong. Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature*, 480(7377):336343, December 2011.

[51] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, September 2007.

[52] J.g. Morel, M.c. Bokossa, and N.k. Neerchal. Small sample correction for the variance of GEE estimators. *Biometrical Journal*, 45(4):395409, 2003.

[53] G. R. Nakamura, D. P. Fonseca, S. M. O'Rourke, A. L. Vollrath, and P. W. Berman. Monoclonal antibodies to the v2 domain of MN-rgp120: fine mapping of epitopes and inhibition of 47 binding. *PLoS ONE*, 7(6):e39045, 2012.

[54] David C. Nickle, Laura Heath, Mark A. Jensen, Peter B. Gilbert, James I. Mullins, and Sergei L. Kosakovsky Pond. HIV-Specific probabilistic models of protein evolution. *PLoS ONE*, 2(6):e503, June 2007.

[55] Morten Nielsen, Claus Lundegaard, Peder Worning, Sanne Lise Lauemller, Kasper Lamberth, Sren Buus, Sren Brunak, and Ole Lund. Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. *Protein Science*, 12(5):10071017, 2003.

[56] Gilad Ofek, F. Javier Guenaga, William R. Schief, Jeff Skinner, David Baker, Richard Wyatt, and Peter D. Kwong. Elicitation of structure-specific antibodies by epitope scaffolds. *Proceedings of the National Academy of Sciences*, September 2010.

[57] Ab Osterhaus, Ron Fouchier, and Guus Rimmelzwaan. Towards universal influenza vaccines? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 366(1579):2766–2773, October 2011. PMID: 21893539.

[58] Wei Pan. On the robust variance estimator in generalised estimating equations. *Biometrika*, 88(3):901–906, September 2001. ArticleType: research-article / Full publication date: Sep., 2001 / Copyright 2001 Biometrika Trust.

[59] Wei Pan and Melanie M. Wall. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine*, 21(10):14291441, 2002.

[60] A. S. Perelson and G. F. Oster. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J. Theor. Biol.*, 81(4):645670, December 1979.

[61] A. Pinter, W. J. Honnen, Y. He, M. K. Gorny, S. Zolla-Pazner, and S. C. Kayman. The V1/V2 domain of gp120 is a global regulator of the sensitivity of primary human immunodeficiency virus type 1 isolates to neutralization by antibodies commonly induced upon infection. *J. Virol.*, 78(10):52055215, May 2004.

[62] A. Pinter, W. J. Honnen, S. C. Kayman, O. Trochev, and Z. Wu. Potent neutralization of primary HIV-1 isolates by antibodies directed against epitopes present in the V1/V2 domain of HIV-1 gp120. *Vaccine*, 16(19):18031811, November 1998.

[63] Punnee Pitisuttithum, Peter Gilbert, Marc Gurwith, William Heyward, Michael Martin, Fritz vanGriensven, Dale Hu, and Jordan W. Tappero. Randomized, DoubleBlind,

PlaceboControlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV1 vaccine among injection drug users in bangkok, thailand. *Journal of Infectious Diseases*, 194(12):1661–1671, December 2006.

[64] Stanley A. Plotkin. Correlates of protection induced by vaccination. *Clinical and Vaccine Immunology*, 17(7):1055–1065, July 2010.

[65] Stanley A. Plotkin and Peter B. Gilbert. Nomenclature for immune correlates of protection after vaccination. *Clinical Infectious Diseases*, 54(11):1615–1617, June 2012.

[66] Li Qin, Peter B. Gilbert, Lawrence Corey, M. Juliana McElrath, and Steven G. Self. A framework for assessing immunological correlates of protection in vaccine trials. *Journal of Infectious Diseases*, 196(9):1304–1312, November 2007.

[67] Supachai Rerks-Ngarm, Punnee Pitisuttithum, Sorachai Nitayaphan, Jaranit Kaewkungwal, Joseph Chiu, Robert Paris, Nakorn Premsri, Chawetsan Namwat, Mark de Souza, Elizabeth Adams, Michael Benenson, Sanjay Gurunathan, Jim Tartaglia, John G. McNeil, Donald P. Francis, Donald Stablein, Deborah L. Birx, Supamit Chunsuttiwat, Chirasak Khamboonruang, Prasert Thongcharoen, Merlin L. Robb, Nelson L. Michael, Prayura Kunasol, and Jerome H. Kim. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in thailand. *New England Journal of Medicine*, 361(23):2209–2220, December 2009.

[68] Douglas D. Richman, Terri Wrin, Susan J. Little, and Christos J. Petropoulos. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):4144–4149, April 2003. PMID: 12644702 PMCID: PMC153062.

[69] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, March 1995. ArticleType: research-article / Full publication date: Mar., 1995 / Copyright 1995 American Statistical Association.

[70] M Rolland, PT Edlefsen, BB Larsen, S Tovanabutra, E Sanders-Buell, T Hertz, AC deCamp, CA Magaret, H Ahmed, M Juraska, H Liao, M Bonsignori, K Hwang, L Chen, P Konopa, S Nariya, JN Stoddard, K Wong, H Zhao, W Deng, BS Maust, M Bose, S Howell, A Bates, M Lazzaro, A OSullivan, E Lei, A Bradfield, G Ibitanumo, V Assawadarachai, RJ OConnell, MS deSouza, S Nitayaphan, S Rerks-Ngarm, ML Robb, C Carrico, S Menis, W Schief, JS McLellan, I Georgiev, PD Kwong, BF Haynes, NL Michael, PB Gilbert, JI Mullins, and Kim JH. Increased HIV-1 vaccine efficacy against viruses with genetic signatures in env- V1/V2 revealed by sieve analysis. *Nature*, 2012.

[71] Morgane Rolland, Sodsai Tovanabutra, Allan C. deCamp, Nicole Frahm, Peter B. Gilbert, Eric Sanders-Buell, Laura Heath, Craig A. Magaret, Meera Bose, Andrea Bradfield, Annemarie O'Sullivan, Jacqueline Crossler, Teresa Jones, Marty Nau, Kim Wong, Hong Zhao, Dana N. Raugi, Stephanie Sorensen, Julia N. Stoddard, Brandon S. Maust, Wenjie Deng, John Hural, Sheri Dubey, Nelson L. Michael, John Shiver, Lawrence Corey, Fusheng Li, Steve G. Self, Jerome Kim, Susan Buchbinder, Danilo R. Casimiro, Michael N. Robertson, Ann Duerr, M. Juliana McElrath, Francine E. McCutchan, and James I. Mullins. Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nature Medicine*, 17(3):366–371, 2011.

[72] Joseph P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692, September 1990. ArticleType: research-article / Full publication date: Sep., 1990 / Copyright 1990 American Statistical Association.

[73] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, March 2005.

[74] Sampa Santra, Hua-Xin Liao, Ruijin Zhang, Mark Muldoon, Sydeaka Watson, Will Fischer, James Theiler, James Szinger, Harikrishnan Balachandran, Adam Buzby, David Quinn, Robert J. Parks, Chun-Yen Tsao, Angela Carville, Keith G. Mansfield, George N. Pavlakis, Barbara K. Felber, Barton F. Haynes, Bette T. Korber, and Norman L. Letvin. Mosaic vaccines elicit CD8+ t lymphocyte responses in monkeys that confer enhanced immune coverage of diverse HIV strains. *Nature medicine*, 16(3):324–328, March 2010. PMID: 20173754 PMCID: PMC2834806.

[75] Michael S Seaman, Holly Janes, Natalie Hawkins, Lauren E Grandpre, Colleen Devoy, Ayush Giri, Rory T Coffey, Linda Harris, Blake Wood, Marcus G Daniels, Tanmoy Bhattacharya, Alan Lapedes, Victoria R Polonis, Francine E McCutchan, Peter B Gilbert, Steve G Self, Bette T Korber, David C Montefiori, and John R Mascola. Tiered categorization of a diverse panel of HIV-1 env pseudoviruses for assessment of neutralizing antibodies. *Journal of virology*, 84(3):1439–1452, February 2010. PMID: 19939925.

[76] Derek J. Smith, Alan S. Lapedes, Jan C. de Jong, Theo M. Bestebroer, Guus F. Rimmelzwaan, Albert D. M. E. Osterhaus, and Ron A. M. Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, July 2004.

[77] Stefan Tenzer, Edmund Wee, Anne Burgevin, Guillaume Stewart-Jones, Lone Friis, Kasper Lamberth, Chih-hao Chang, Mikkel Harndahl, Mirjana Weimershaus, Jan Gerstoft, Nadja Akkad, Paul Klenerman, Lars Fugger, E. Yvonne Jones, Andrew J. McMichael, Sren Buus, Hansjrg Schild, Peter van Endert, and Astrid K. N. Iversen. Antigen processing influences HIV-specific cytotoxic t lymphocyte immunodominance. *Nature Immunology*, 10(6):636–646, June 2009.

[78] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, January 1996. ArticleType: research-article / Full publication date: 1996 / Copyright 1996 Royal Statistical Society.

[79] Mark J van der Laan, Sandrine Dudoit, and Katherine S Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical applications in genetics and molecular biology*, 3:Article15, 2004. PMID: 16646793.

[80] Laura M. Walker, Michael Huber, Katie J. Doores, Emilia Falkowska, Robert Pejchal, Jean-Philippe Julien, Sheng-Kai Wang, Alejandra Ramos, Po-Ying Chan-Hui, Matthew Moyle, Jennifer L. Mitcham, Phillip W. Hammond, Ole A. Olsen, Pham Phung, Steven Fling, Chi-Huey Wong, Sanjay Phogat, Terri Wrin, Melissa D. Simek, Protocol G. Principal Investigators, Wayne C. Koff, Ian A. Wilson, Dennis R. Burton, and Pascal Poignard. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature*, 477(7365):466–470, September 2011.

[81] Laura M. Walker, Sanjay K. Phogat, Po-Ying Chan-Hui, Denise Wagner, Pham Phung, Julie L. Goss, Terri Wrin, Melissa D. Simek, Steven Fling, Jennifer L. Mitcham, Jennifer K. Lehrman, Frances H. Priddy, Ole A. Olsen, Steven M. Frey, Phillip W. Hammond, Stephen Kaminsky, Timothy Zamb, Matthew Moyle, Wayne C. Koff, Pascal Poignard, and Dennis R. Burton. Broad and potent neutralizing antibodies from an african donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289, October 2009.

[82] David I Warton. Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics*, 67(1):116–123, March 2011. PMID: 20528857.

[83] Xiping Wei, Julie M Decker, Shuyi Wang, Huxiong Hui, John C Kappes, Xiaoyun Wu, Jesus F Salazar-Gonzalez, Maria G Salazar, J Michael Kilby, Michael S Saag, Natalia L Komarova, Martin A Nowak, Beatrice H Hahn, Peter D Kwong, and George M Shaw. Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307–312, March 2003. PMID: 12646921.

[84] John M Williamson, Somnath Datta, and Glen A Satten. Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59(1):36–42, March 2003. PMID: 12762439.

[85] J.H. Won and Seung-Jean Kim. Maximum likelihood covariance estimation with a condition number constraint. In *Fortieth Asilomar Conference on Signals, Systems and Computers, 2006. ACSSC '06*, pages 1445–1449, 2006.

[86] Xuesong Yu, Peter B. Gilbert, Catarina E. Hioe, Susan Zolla-Pazner, and Steven G. Self. Statistical approaches to analyzing HIV-1 neutralizing antibody assay data. *Statistics in Biopharmaceutical Research*, 4(1):1–13, January 2012.

[87] S L Zeger and K Y Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, March 1986. PMID: 3719049.

[88] Andreas Ziegler. *Generalized estimating equations*. Springer, New York, 2011.

[89] S. Zolla-Pazner, T. Cardozo, A. deCamp, B. Haynes, J. Kim, X Kong, N. Michael, S. Rerks-Ngarm, and C. Williams. V2-reactive antibodies in RV144 vaccinees' plasma. In *AIDS Vaccine*, Bangkok, Thailand, 2011.

[90] Susan Zolla-Pazner and Timothy Cardozo. Structurefunction relationships of HIV-1 envelope sequence-variable regions refocus vaccine design. *Nature Reviews Immunology*, 10(7):527–535, July 2010.